The Nature of Intention

Gil Alexander Percival

University College London

Department of Philosophy

PhD

Declaration

I, Gil Alexander Percival, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

G.A.P.

September 2013

Abstract

Imagine you face the following choice: either spending the evening at a party, or going to the library and continuing with the paper you have been working on. You have been working hard recently and have a strong desire to go to the party. On the other hand, you have an important deadline coming up and need to make progress with the paper. Whichever way you decide, once the decision is made you will enter into new kind of state, adopting a particular kind of attitude towards your own future. This is the state of intention. What is the nature of this state? The thesis to be defended over the following five chapters is that intention is a primitive and irreducible mental state, non-analyzable in terms of any other, supposedly more basic, folk-psychological states or attitudes, or combination thereof, such as desire and belief. I make two important claims about intention. One is that intention is a state that, like belief, has an aim. However, whereas the aim of belief is knowledge, the aim of intention is self-control, or determining what one will do in the future. I argue that it is the fact that intention aims at self-control that explains certain distinctive normative features of intention that distinguish intention from desire and belief. The other claim is that intention is a kind of disposition – the disposition of an agent to pursue an aim or goal. I argue that this explains certain distinctive causal and descriptive features of intention that distinguish it from desire and belief.

Table of Contents

Abstract 3
Introduction 6
Chapter One 10
Bratman on the Irreducibility of Intention 10
1. Introduction 10
1.1 The Planning Theory 10
1.2 Objections to the Predominant Desire Theory 15
1.3 Objections to the Supplemented Desire Theory 21
1.4 Intention and Planning 30
1.5 Conclusion 33
Chapter Two 36
Cognitivism and the Requirements of Practical Rationality 36
2. Introduction 36
2.1 The Cognitivist Explanation of the Requirements 37
2.2 Bratman's Objections 44
2.3. Infallibility 47
2.4 Further Objections to Infallibility 53
2.5 Conclusion 59
Chapter 3 60
Congitivism and the Epistemic Conditions on Intentional Action 60
3. Introduction 60
3.1 Velleman's theory 61
3.2 Intending and Deciding to Believe 68
3.3 Can't Epiphenomenalists Form Intentions? 73
3.4 Setiya's Theory 76
3.5 Anscombe on Practical Knowledge 84
3.6 Conclusion 90
Chapter 4 91
Intention, Belief and the Normative Role of Knowledge 91
4. Introduction 91
4.1 Two Types of Cognitivist Theory 92
4.2 The Fundamental Norm of Belief 93
4.3 Williamson's Argument for the Knowledge View 97
4.4 Further Arguments for the Knowledge View 103

4.5 Objections to Weak and Strong Cognitivism 105
4.6 Potential Replies to the Objections 109
4.7 Conclusion 117
Chapter 5 118
Intentional Action and Causal Deviance 118
5. Introduction 118
5.1 Bishop and the Problem of Causal Deviance 122
5.2 Hyman's Solution to the Problem of Causal Deviance 135
5.3 Are Dispositions Causally Relevant? 145
5.4 Two Further Objections to the Causal Theory 155
5.5 Conclusion 158
Acknowledgements 159
Bibliography 160

Introduction

Imagine you face the following choice: either spending the evening at a party you have been invited to, or going to the library and continuing with the paper you have been working on. You have been working hard recently and have a strong desire to go to the party. On the other hand, you have an important deadline coming up and need to make progress with the paper. Whichever way you decide, once the decision is made you will enter into new kind of state, adopting a particular kind of attitude towards your own future. This is the state of intention. What is the nature of this state? The thesis to be defended over the following five chapters is that intention is a primitive and irreducible mental state, which is non-analyzable in terms of any other, supposedly more basic, folk-psychological states or attitudes, or combination thereof, such as desire and belief.

Though my aim is to defend the irreducibility of intention, I do not adopt a position of quietism about intention. I do believe that there are interesting things to be said about what intentions are. In this respect, there is a parallel with certain claims that Timothy Williamson (2000) makes about the nature of knowledge. According to Williamson, knowledge is a basic mental state, rather than something to be analyzed in terms of belief, truth and certain other conditions. However, for Williamson, this does not mean that there is nothing interesting to say about knowledge, or that we cannot use the concept to shed light on the nature of certain other concepts, such as evidence, evidential probability and assertion.

Over the next five chapters, I make two central claims about intention. One is that intention is a state that, like belief, has an aim. However, whereas the aim of belief is knowledge, the aim of intention is self-control, or determining what one will do in the future. I argue that it is the fact that intention aims at self-control that explains certain distinctive normative features of intention, to be introduced in chapter one. The other claim that I make is that intention is a kind of disposition – the disposition of an agent to pursue an aim or goal. Just as Williamson uses the primitiveness of knowledge to elucidate certain other concepts, I believe that we can use the idea that intention is a primitive kind of dispositional state to shed light on the concept of intentional action. Intentional action is not to be analyzed in terms of an event, plus an intention, plus certain further conditions specifying the appropriate event-causal relation between the

two. Intentional action just is the manifestation of intention. I argue that this explains certain distinctive causal and descriptive features of intention, also to be introduced in chapter one.

Over the first four chapters, I examine two prominent traditions in the literature on intention. One is Michael Bratman's (1987) 'planning theory' of intention. The other is what Bratman (1991) refers to as 'cognitivism' about intention. Bratman associates the term cognitivism with two different ideas: first, that intending to act necessarily involves believing that one will so act. Second, that the rational demands or requirements that intentions be consistent and means-end coherent are derived from corresponding rational requirements on involved beliefs. Generally, I will be using the term 'cognitivist' in a less restricted way so as to refer to anyone who holds the first of these ideas - that intending to act is constitutively tied to believing that one will so act.

In chapter one I examine Bratman's planning theory. Bratman makes two central claims. One is that intention has certain distinctive descriptive and normative characteristics, which cannot be adequately accounted for if we analyse an intention to act either as the predominant desire to act, or the combination of the predominant desire to act plus the belief that one will so act because one predominantly desires to. The other claim is that intentions are metaphysically bound up with planning agency – they are the atoms or building blocks of larger plans. It is the fact that intentions are the atoms or building blocks of larger plans that explains the distinctive features of intention. In chapter one I agree with the first of these claims. However, I reject the second. I propose that intentions are a distinctive type of mental state the aim or point of which is self-control, or determining one's future actions, as an alternative explanation of the normative features of intention.

In chapter two I turn to cognitivism about intention. In chapter one I reject one potential cognitivist analysis of intention – the view that an intention to act consists in a predominant desire to act plus the belief that one will so act because one predominantly desires to. However, there are number of other cognitivist proposals in the literature still to be considered. For example, some philosophers, such as Gilbert Harman (1976) and Kieran Setiya (2007), argue in different ways that an intention to ϕ *just is* a special kind of belief that one will ϕ. Another proposal defended in the literature is that intending to ϕ

constitutively involves predominantly desiring to ϕ because one believes that one will ϕ, rather than vice versa. This is a rough approximation of the account defended by David Velleman (1989). A full defence of the irreducibility of intention will have to provide adequate reasons for rejecting all these proposals as well.

Over the course of chapters two and three I consider two principal motivations for cognitivism. One is the idea that cognitivism can explain the rational demands of intention-consistency and means-end coherence governing intention by appealing to corresponding demands for consistency and coherence on the purportedly involved beliefs. The other is the idea, defended by Velleman (1989) and Setiya (2007), that analyzing intention as identical with or as constitutively involving some sort of causally self-fulfilling belief can explain certain ideas, inspired by Elizabeth Anscombe (1957), concerning the epistemic conditions on intentional action. In chapter two I argue that, contrary to Bratman (2009), who raises certain problems for the cognitivist derivation of the rational demands for consistency and means-end coherence of intentions, it might be possible to explain the requirements governing intention by appealing to corresponding demands on purportedly involved beliefs. Nevertheless, this is not a decisive reason in favour of cognitivism. There are other possible explanations of the requirements on intention – for instance, that the aim or point of forming intentions is self-control. In chapter three I argue that Velleman's and Setiya's strategy from explaining the proposed epistemic conditions on intentional action is problematic.

In chapter four I present objections to cognitivism of a more general nature. There is nothing necessarily wrong or objectionable about intending to do something that one does not know one will do. However, defending Williamson's (2000) claim that the fundamental norm of belief is knowledge, I argue that it follows implausibly from cognitivism that there is necessarily something wrong with this. In essence, my argument is that intending to act cannot constitutively involve believing that one will act because intention and belief have fundamentally different normative properties. In forming beliefs we aspire to know. However, in forming intentions we do not aspire to know how we will act in the future, but to determine or control how we will act in the future.

In chapter five I turn to the relation between intention and intentional action. I defend the causal theory of intentional action, according to which an event is intentional of an

agent if and only if it is caused by an intention of that agent with an appropriate content. A long-standing objection to the causal theory of intentional action is the problem of so-called deviant causal chains. I argue that in order to solve the problem of deviant causal chains we should understand the relation between intention and intentional action, not in event-causal terms, but as the manifestation or exercise of a certain type of disposition – the disposition of an agent to pursue an aim or goal. I argue that aside from solving the problem of deviant causal chains, as well as fitting comfortably with the idea that the aim of intention is self-control, the dispositional analysis of intention can also account for the various descriptive or causal features of intention highlighted by Bratman. It can also help settle certain additional worries some may have about the causal theory of intentional action, aside from the problem of causal deviance.

Chapter One

Bratman on the Irreducibility of Intention

1. Introduction

In his *Intention, Plans and Practical Reason* (1987), Michael Bratman develops what he calls a 'planning theory of intention'. Bratman's planning theory makes two important claims about intentions. One is that intentions are distinct and irreducible states that are non-analyzable in terms of either desire or some combination of desire and belief. The other is that intentions are metaphysically bound up with planning agency and with a more general theory of bounded rationality. Bratman's argument against desire-belief reductionism involves two steps. Firstly, Bratman argues that intentions are associated with certain distinctive characteristics related to practical reasoning and action. Secondly, he argues that these characteristics cannot be accounted for if we understand intention either as mere 'predominant desire', or, alternatively, as the combination of a predominant desire to act plus the belief that one will act because one predominantly desires to. In this chapter I examine Bratman's argument for his planning theory of intention. I agree with Bratman that the distinctive characteristics of intention cannot be accounted for if we understand intention either as mere 'predominant desire' (section 1.2), or as the combination of a predominant desire to act plus the belief that one will act because one predominantly desires to (section 1.3). However, I reject Bratman's claim that intentions are metaphysically bound up with planning agency (section 1.4).

1.1 The Planning Theory

In his *Intention, Plans and Practical Reason* (1987), Michael Bratman develops what he calls a 'planning theory of intention'. In setting out his theory, Bratman notes that the term 'intention' is often used in different senses. Sometimes it is used to characterise an agent's actions. At other times it is used to characterise her state of mind (1987: 2). The distinction is that between 'intention *in* action' and 'intention *to* act'[1]. The former notion

[1] The first person to distinguish between the different uses of the concept of intention was G.E.M. Anscombe (1957: 1). One difference between Anscombe's formulations of the distinct applications of the concept and Bratman's is that, whereas Bratman speaks of intending to act (e.g. intending this morning to

of intention in action encompasses talk of both acting *intentionally* and acting *with an intention*. The latter notion of intending to act encompasses talk of intending to do something, either beginning now (what Bratman refers to as 'present-directed' intention to act) or at a later time (what Bratman calls 'future-directed' intention to act). Bratman's approach to thinking about intention involves treating as central the case of future-directed intention (1987: 3-4). His view is that we must first try to understand what a future-directed intention is. Light will then be shed on the other senses of intention. Bratman argues that in order justify this approach of treating future-directed intention as central some explanation should be offered of why we form future-directed intentions in the first place and why forming future-directed intentions is so important to our lives. Bratman's answer to the question of why we form future-directed intentions, rather than always just crossing our bridges when we come to them, centres on the idea that we are 'planning agents'. Bratman argues that we need to form plans, and we form future-directed intentions because these are the components that make up plans. As he puts it, plans are "intentions writ large" (1987: 29).

Bratman outlines two main reasons why forming future-directed intentions, and so by extension forming plans, benefits us and is important. Firstly, he says that forming future-directed intentions facilitates co-ordination both intrapersonally and interpersonally. He says that future-directed intentions do this because they support the expectation that the agent will do the thing she formed the intention to do. He argues that this enables both other people and the agent herself to plan, act and organise their affairs on the basis of that expectation. Secondly, Bratman argues that if we didn't form future-directed intentions the influence of deliberation would only extend as far as the time of deliberation. Forming future-directed intentions enables us to extend the influence of our deliberations beyond the present moment. This second factor is important given our bounded rationality – our limited capacities for deliberation and processing information. The thought here is that it may be better to form an intention to act in advance of the time of action itself because it may be likely that at the time of action one will have less time to deliberate and think through the options, one may be

---

pump the water), Anscombe speaks of "expressions of intention" (2000: 1). Richard Moran and Martin J. Stone (2011) have argued that recognizing the often-overlooked fact that Anscombe frequently speaks, not of intentions to act, but of expressions of intention for the future, is crucial to understanding her views on intention. I discuss Anscombe in more detail in chapter three.

distracted, or one may be susceptible to temptations or other influences that might negatively bias one's judgement.

According to Bratman, in order to fully account for the various roles that intentions play in our planning agency, we need to recognize intention as a distinctive psychological state or attitude non-analysable in terms of other, supposedly more basic psychological states or attitudes. According to the traditional analysis of intentional action, which Bratman terms the 'desire-belief theory of intention in action' (1987: 6), intentional action is behaviour that stands in some appropriate relation to an agent's desires and beliefs. Bratman rejects the idea that we can simply start out with this prior conception of intentional action and then hope to successfully extend this analysis in some shape or form to the case of intentions to act[2]. Bratman describes his strategy for rejecting the reducibility of intention as appealing to functionalism within the philosophy of mind (1987: 9). His argument involves identifying the state of intending to act by 'embedding' it in a "web of regularities and norms" concerning extended practical reasoning and action (1987: 10). In other words, he describes certain descriptive and normative features of the state or attitude of intending relating to dispositions or propensities to reason and to act. He then attempts to differentiate intention from states of desire and belief, or some combination thereof, by showing that any analysis in terms of desire, or desire plus belief, cannot explain these features.

Bratman argues that the various descriptive and normative features associated with the state or attitude of intending, which differentiate intending from anything else, all relate to the fact that intending "involves a characteristic kind of commitment" (1987: 15). There are various ways in which individuals can commit themselves to a course of action,

[2] For a philosopher who understands intentional action as behaviour that stands in an appropriate relation to the agent's desires (or 'pro-attitudes') and beliefs, and who, by a strategy of extension, treats intention as reducible to desire and belief, see Donald Davidson (1963). In his later (1978), Davidson explicitly describes himself in his earlier (1963) paper as viewing intentions as implicit in the agent's rationalizing desires and beliefs. Davidson (1978) subsequently came to view this analysis of intention as problematic. In the introduction to his (1980) collected papers, *Essays on Actions and Events*, he writes that of the three different uses of the concept of intention - acting with an intention, acting intentionally, and intending to act - the notion of intending to act "came to seem the basic notion on which the others depend; and what progress I made with it partially undermines an important theme in Essay 1 [1963] – that 'the intention with which the action was done' does not refer to an entity or state of any kind" (1980: xiii). Instead, he argued that intentions are 'all-out evaluations' that, in some cases, explain and rationalize an agent's intentional actions alongside her desires and beliefs. On his later account, a future-directed intention to ϕ is the all-out judgment that performing an action of the type, ϕ-ing, would be desirable given the rest of the agent's beliefs about the future. For discussion and criticism of Davidson's later account of intention, see Bratman (1985).

imposing constraints on their practical futures. Jon Elster (1979) has argued that given that human beings are imperfectly rational and susceptible to temptation and other kinds of adverse influences on their reasoning and decision-making, sometimes it can be rational for us to undertake various sorts of 'precommitment'. Such precommitment involves cross-temporal manipulation of one's future self by one's past self by either altering the incentives or payoffs of, or by altering the practical feasibility of, some future course of action. For example, one might precommit to saving a certain percentage of one's earnings for retirement by directing the money to a savings account that imposes significant fines for early withdrawals. Or one might precommit to giving up smoking for a period by moving somewhere where there is nowhere to obtain tobacco, thereby narrowing the range of one's future options. In all such cases, methods are employed that make it difficult or unappealing for one's future self to abandon or renege on the intentions of one's earlier self. However, as will be made clear below, Bratman thinks that forming an intention involves a non-manipulative and more everyday or humdrum kind of commitment that is different from these sorts of methods of self-binding. Crucially, he maintains that it is because an agent's intending to ϕ involves her committing to ϕ-ing in this more ordinary, non-manipulative sense that the intention can support the expectation that she will do what she intends, thereby facilitating intrapersonal and interpersonal co-ordination (1987: 17/8).

Bratman argues that the descriptive or causal component of the commitment characteristic involved in future-directed intending has two dimensions, which he refers to as 'volitional' and 'reasoning-centred'. Bratman says that future-directed intentions are 'volitional' because they are 'conduct-controlling pro-attitudes' (1987: 16). In this respect, he says they differ from desires, which are 'conduct-influencing pro-attitudes'. Bratman claims that the difference is that having a desire to do something might *influence* whether or not I do that thing, but it will not determine that I will do it insofar as my desire to do it is merely one reason for doing it to be weighed among other such desire-belief reasons. However, others things beings equal, if I have formed an intention to do something, as long as I do not reconsider the intention at some later point and as long as nothing else interferes, my having formed that intention determines that I will do that thing. The reasoning-centred dimension of commitment is different from the volitional dimension in that, while the volition dimension concerns the disposition of the agent to act, the reasoning-centred dimension concerns the agent's dispositions to reason, or

refrain from further reasoning, between the time of intention-formation and the time of action. The reasoning-centred dimension of commitment has two aspects. Firstly, Bratman says that when we form an intention to do something we 'settle' on doing it. A decision is made and deliberation is discontinued. Bratman expresses this by saying that intention has "a characteristic stability or inertia" (1987: 16), and that retention and non-reconsideration of a prior intention is the "default option" (1987: 17). The basic thought here is that we don't normally spontaneously reopen deliberation and reconsider an intention unless there is some change in the situation that gives us reason to do so. We are disposed not to reconsider unless some such change in the situation arises. Secondly, Bratman argues that when I form an intention to act this normally affects my further reasoning between now and the time of action. In forming the future-directed intention to ϕ, I will then be disposed to reason from prior intentions to further intentions and from ends to means. Furthermore, my formation of the intention to ϕ will constrain what other intentions I subsequently form for the reason that I will be disposed to make all my intentions consistent with each other and with my beliefs.

Bratman also argues that there is a normative component to the commitment characteristic of future-directed intention, in addition to its causal or descriptive features. Bratman distinguishes between two kinds of norms of intention, 'internal' and 'external'. He says that the internal norms of intention govern an agent's practical reasoning whilst treating her background framework of prior plans and intentions as fixed. These internal norms govern deliberation from within the agent's own 'plan-constrained' perspective. By contrast, he says that the function of the external norms of intention is in reaching an overall assessment of an agent's behaviour, either by the agent herself or by some third party, that is unconstrained by the agent's background of prior plans and intentions. Bratman's views on the external norms governing intention, which concern when it is rational to form an intention taking into account the rationality of the agent's background of prior intentions and plans formed at some earlier time, and when it is rational to reconsider or not to reconsider a prior intention and what constitute good habits of reconsideration and non-reconsideration, are complex and extend across a number of chapters. A full discussion of them would take me too far afield for the purposes of this chapter. Instead I will only consider what Bratman calls the internal norms of intention. These are 'means-end coherence' and the 'consistency constraints' on intention (1987: 31). Means-end coherence is the demand that the agent reason from

prior intentions to further intentions and from ends to means in a manner that is sufficient for achieving her ends. Bratman distinguishes between two types of consistency constraint on intention. He says that an agent's intentions should be 'internally consistent' in the sense that it should be possible for them to be collectively realized by the agent. Further, he says that an agent's intentions should be 'strongly consistent', relative to the agent's beliefs. This means that it should be possible for an agent to realise her intentions assuming that the world is as she believes it to be (1987: 31).

Having outlined Bratman's account of the distinctive marks or characteristics of intention, we are now in a position to examine his arguments for the non-reducibility of intention either to a predominant desire to act, or a predominant desire plus the belief that one will so act because one predominantly desires to.

1.2 Objections to the Predominant Desire Theory

In rejecting the claim that future-directed intention is reducible to some complex of desire and belief, Bratman anticipates in two stages how such a reduction might work. In the first stage, Bratman proposes that the reductionist about future-directed intention might treat future-directed intentions as 'predominant desires'. According to Bratman's definition of predominant desire, having the predominant desire to ϕ involves desiring to ϕ more than any other option that one believes to be incompatible with ϕ-ing (1987: 18). Bratman finds the identification of intentions with predominant desires problematic for various reasons. In the second stage, he suggests that the reductionist might supplement the predominant desire analysis with a belief condition according to which intention necessarily or constitutively involves the belief that one will act as one predominantly desires to act. Bratman finds this problematic as well. In this section I focus on Bratman's rejection of the identification of intention with predominant desire. I will agree with Bratman that we should reject the identification of intention with predominant desire, and I will also offer some supplementary objections to that analysis. In the next section I turn to Bratman's rejection of a supplemented version of the predominant desire analysis.

Bratman raises three objections to the identification of intention with predominant desire. His first objection is that the analysis of intention in terms of predominant desire will not be able to account for the reasoning-centred dimension of commitment. He says that having a predominant desire to ϕ does not ensure that the agent sees the question of whether to ϕ as settled. He also argues that merely having a predominant desire to ϕ will not dispose the agent to reason from the intention to ϕ to further intentions, such as to means of ϕ-ing. Bratman gives the example of choosing between attending a concert and going to the library (1987: 18). Supposing that I predominantly desire to go the library, Bratman says that this alone does not guarantee that I see the matter as settled. Moreover, he argues that if I am still treating it is an open question whether to go the library or the concert I am unlikely to be disposed to figure out and intend means to going to the library.

Bratman's second objection to the predominant desire analysis of intention is that it cannot account for the volitional dimension of commitment. As already mentioned, Bratman describes intention as a 'conduct-controlling pro-attitude'. However, he says that desires are merely 'conduct-influencing'. Returning to the example of the choice between going to the library or attending the concert, Bratman argues that if I still see the question of what to do as open, and so I am not yet disposed to reason towards and intend the relevant means to going to the library, then it is unclear how my possessing the predominant desire to go to the library at noon could by itself determine what I am going to do. While my possessing the predominant desire to go to the library at noon might incline me towards ultimately settling in favour of this option, it does not seem sufficient to determine my actions.

Thirdly and finally, Bratman argues that while it is rationally criticisable to form intentions that one believes are incompatible, it is perfectly *possible* to form intentions that one believes are incompatible. However, Bratman objects that if intentions were predominant desires then forming intentions that one believes are incompatible would be impossible. To see why Bratman says this, recall that, as he defines it, the concept of predominant desire has a belief component built into it. According to Bratman, predominantly desiring to ϕ necessarily involves desiring to ϕ more than anything that one *believes* is incompatible with ϕ-ing. This entails that for any two options that one believes are incompatible, one cannot predominantly desire to do both. If intentions are

predominant desires, it would follow that one cannot intend to do any two things that one believes are incompatible. Bratman claims that this is false. He thinks that a person can intends to do things that she believes are incompatible, though he also thinks that this would be irrational.

Bratman's third objection could be met with two responses. Firstly, the defender of the predominant desire analysis might try to weaken Bratman's definition of predominant desire as follows:

> *(PD*)* For any agent S, S predominantly desires to ϕ if and only if S desires to ϕ no less than anything S believes to be inconsistent with ϕ-ing.

(PD*) has it built into it that subjects can simultaneously have incompatible predominant desires. However, there are two problems with (PD*). Firstly, the motivation for introducing the notion of predominant desire into an analysis of intention in the first place would be to express the idea that intending to do something in some sense involves desiring to do that thing the most, or more than anything else. But with (PD*) this sense is lost. Secondly, (PD*) would have the peculiar consequence that an agent with no desires would intend to do everything. For any value for ϕ, an agent with no desires would desire to ϕ no less than anything she believes to be incompatible with ϕ-ing. Therefore, if intention were identical with predominant desire, it would follow that she intends to ϕ.

According to the second possible response to Bratman's third objection, the defender of the predominant desire analysis could reject Bratman's claim that it is possible to form intentions that one believes are incompatible. Suppose one believes that two options, say, finishing a paper and going out to dinner, are incompatible with one another. Further, suppose that one possesses the prior intention to finish the paper. It might be argued that if one genuinely considers the question of whether to go out for dinner then, given the current state of one's beliefs, one would automatically re-open the question of whether to finish the paper. In which case, one is no longer settled in favour of finishing the paper. The reductionist could acknowledge that it is possible to form intentions – such as the intention to finish the paper and the intention to go out for dinner - that one *did* believe were inconsistent, having now forgotten this fact, or that one *would* believe

were inconsistent, if one took the time to really think about it. However, this is not the same as forming intentions *whilst* believing they are incompatible. Forgetting that one's intentions are incompatible is not the same as believing it, but merely not occurrently thinking about it. Forgetting implies loss, even if only temporarily, of belief. Failing to see or to realize that *p* also implies the absence of belief that *p*. The reductionist might argue that as soon as one remembers or comes to see the inconsistency of one's intentions, and so regains or acquires the belief that they are incompatible, one would immediately be faced with the task of settling anew the question of what to do. It seems to me that this is not an implausible response to Bratman's third objection.

Irrespective of the strength of Bratman's third objection, I think his first and second objections are convincing. Is there a more sophisticated analysis of intention in terms of predominant desire that can avoid these first and second objections? Michael Ridge (1998) has proposed the following account:

"A intends to ϕ if and only if (a) A has a desire to ϕ, (b) A does not believe that ϕ-ing is beyond her control, (c) A's desire to ϕ is a *predominant* one, which is just to say that there is no desire to ψ, such that A does not believe ψ-ing is beyond her control, she desires to ψ as much as or more than she desires to ϕ, and she believes that a necessary means to her ϕ-ing is that she refrains from ψ-ing, (d) A has a desire not to deliberate any more about whether to ϕ unless new, relevant information comes to light" (1998: 163)

Ridge views desires as dispositions to pursue what is desired, including by taking whatever one believes to be means[3]. He incorporates condition (d) in order to meet Bratman's objection that the analysis of intention in terms of predominant desire cannot account for the stability or inertia of intention. However, even if a more sophisticated analysis of intention in terms of predominant desire, along the lines of Ridge's account, could offer a satisfactory response to Bratman's first and second objections, there are further reasons for rejecting the predominant desire analysis of intention that are clearly problems for even Ridge's account. The first is suggested by Richard Holton (2009: 13) and relates to 'Buridan' cases. Burdian cases are named after the example, attributed to Jean Buridan, of an ass positioned between two equidistant and equally attractive bales of hay, which starves to death because she is unable to choose between them. Buridan

[3] As will be seen in chapter five, in this respect there is a similarity between Ridge's views and the views of John Hyman (*forthcoming*). Both think intentional actions are caused by desires, which they argue are dispositions to pursue an aim or goal.

cases are examples in which an agent faces a choice between two or more options that she is indifferent between or equally drawn towards, either because she regards the reasons in favour of the different options as commensurable but equally strong, or because she regards the reasons in favour of them as incommensurable and immeasurable on the same scale of value. We seem to face such choices on a regular basis. Consider, for instance, choosing between two similar items in a supermarket, or between two books in a bookstore that one is equally keen to read, or taking a walk through the woods and choosing between two equally appealing paths. However, in Buridan cases we do not normally remain paralyzed or trapped in an impasse. Eventually we form the intention to do one thing or the other. The predominant desire analysis of intention looks hard-pressed to explain this fact. If, following Bratman, we define the predominant desire to ϕ as desiring to ϕ more than any other option that one believes to be incompatible with ϕ-ing, then in Burdan cases we do not predominantly desire either option. Supposing that intentions are predominant desires, it would follow that we could not intend to do any of the options in Buridan cases. On the other hand, if, in accordance with (PD*), we define the predominant desire to ϕ as desiring to ϕ no less than anything that one believes to be inconsistent with ϕ-ing, then in Buridan cases one would predominantly desire all of the options. After all, one is drawn towards each no less than any other. Supposing that intentions are predominant desires, then it would follow implausibly that in Buridan cases one irrationally intends to do all of the options.

Donald Davidson (1985) argues that in Buridan cases, where an agent is unable to find intrinsic grounds for settling in favour of one of the options over all of the others, she will instead seek extrinsic grounds, such as the result of a coin-toss. However, granting this is sometimes the case, it is still not clear how such extrinsic grounds for favouring one of the options could make that option more *desirable* in the eyes of the agent. With respect to Davidson's example of the coin-toss, such a device could only be effective in light of the agent's prior intention to choose according to the results of the toss. Thus the predominant desire theorist will have to say that one of the options becomes more desirable because the agent desires most to do what the coin toss determines. However, this does not sound right at all. The arbitrary results of a coin-toss, or some other random decision-making device, cannot suddenly make one option look more desirable, even if the agent has made it her policy to do what the coin-toss determines. In fact, coin tosses are often resorted to in the hope that they will reveal to us our true

preferences by means of the disappointment we may unexpectedly feel when the results of the toss go a certain way. Furthermore, it is just not correct that in all Buridan cases we resort to such extrinsic grounds for favouring one option over the others. Often we just make a decision one way or the other. When faced with Buridan situations, we often do not have the time or resources to spend endlessly deliberating about what choice to make. In this respect, resolving Buridan-type situations is an important function of forming intentions, and a significant reason for treating intention as distinct from predominant desire.

A second reason for rejecting the predominant desire analysis of intention, also not considered by Bratman, is suggested by David Charles (1989). Charles agrees with Bratman that an important feature of the commitment-characteristic of future-directed intention is that intention has a certain 'inertia', in the sense that one is disposed not to reopen deliberation and reconsider one's prior intentions. He also agrees with Bratman that there are certain norms pertaining to the non-reconsideration of a future-directed intention to ϕ at *t2*. Charles writes that,

"If one intends now to ϕ at t2 it seems required that…One is disposed not to reconsider the options unless, e.g., new evidence is available to one about one's chances of achieving one's goal by a given route or one comes to have a new valuation of that goal (evidence of kind K emerges)…" (1989: 40/1)

Charles argues that the fact that there are rational or normative demands pertaining to the non-reconsideration of one's intention to ϕ at *t2* makes the intention different from desiring most to ϕ at *t2*. He argues that when one forms the intention to do something in the future one is making a kind of commitment to do that thing unless evidence of kind *K* emerges. If one changes one's mind about what to do at *t2* without evidence of kind *K* emerging then there are two possibilities. Either one judges that one's earlier decision was mistaken, in which case it seems rational for one to reconsider one's intention, or one does not judge that one's earlier decision was mistaken, in which case one seems open to rational criticism for undermining what, by one's own lights, was the perfectly rational decision of one's earlier self. However, none of this seems true of desire. One's desires, even one's predominant desires, can change and fluctuate over time without any new evidence or reasons arising to justify that change. For example,

the desire for something sweet can give way to the desire for something savoury. There need be no irrationality in this at all either on the part of one's past or one's future self.

1.3 Objections to the Supplemented Desire Theory

After presenting his objections to the identification of intention with predominant desire, Bratman considers whether these objections might be overcome by supplementing the predominant desire proposal with an additional belief condition. He considers the proposal that the intention to perform an action is identical with the psychological complex of the predominant desire to perform that action, along with the belief that one will perform that action because one predominantly desires to perform it (1987: 19). Bratman raises three objections to the supplemented predominant desire analysis. One worry that he raises is that the additional belief condition does nothing to block his third objection to the identification of intention with predominant desire (1987: 19). According to that objection, the identification of intention with predominant desire would make having intentions that one believes inconsistent impossible. However, for reasons outlined above, it is not clear to me that this objection is ultimately a conclusive one anyway. The defender of the belief-desire analysis might argue that having intentions that one believes are incompatible is indeed impossible.

A second worry that Bratman raises is that the addition of a further belief condition still does not guarantee that in forming the intention to ϕ the agent will see the question of whether to ϕ as settled (1987: 19). Bratman attempts to support this second worry using two examples. The first is as follows:

> "Suppose that I presently have a predominant desire to go to Tanner and (knowing my own work habits) expect that as a result of this desire I will go to Tanner. Could I nevertheless continue to be disposed to deliberate about whether to take the afternoon off? Suppose that I suspect that my predominant desire to go to Tanner is a result of my workaholic tendencies and want to reflect further on the matter. When I step back and try to make a prediction about what I will do, I continue to expect that I will end up going to Tanner. Still, I am suspicious of my motivation and want to think about it some more. Am I settled on going to Tanner, in the sense of being settled that is involved in reasoning-centred commitment? I am inclined to say no" (1987: 19/20).

It seems to me that intuitions about this example could go different ways. There are two ways that the reductionist could interpret this example. On the one hand, if the agent has a predominant desire to go to Tanner, but is suspicious of her motivation and wants to think more about it, then it may be that she is still considering the question of whether to go and has not yet made up her mind one way or the other. In this case, the reductionist could insist that the agent does not believe that she will go to Tanner. The reductionist might concede that given the agent's knowledge of her workaholic tendencies, habits and past patterns of behaviour she may believe it probable, or have some degree of confidence, that she will end up going to Tanner. Thus perhaps if the agent were offered a bet at favourable odds about whether she would end up going to Tanner she would accept it. However, this is consistent with claiming that she does not *flat-out* or *outright* believe that she will go to Tanner. Mere 'expectation' is not necessarily the same as outright belief. To merely expect that p simply involves having some degree of confidence that *p*, or believing it to some extent probable that *p*. However, to flat-out believe that *p* involves being willing to treat *p* as certain, and to take it for granted in one's reasoning and one's actions[4]. On the other hand, the reductionist might argue that if the agent both predominantly desires to go to Tanner and does fully believe that she will go because this is what she predominantly desires to do then she does indeed intend to go Tanner. The fact that she is thinking about her motivations and reflecting on her workaholic tendencies is not necessarily inconsistent with her having this intention. It is possible to reflect on one's motivations for intending to act a certain way without actually reopening deliberation.

The second example Bratman offers is as follows:

> "Suppose I have a fleeting craving for a chocolate bar, one which induces a fleetingly predominant desire to eat one for desert. And suppose that just as fleetingly I notice this desire and judge (in a spirit of resignation, perhaps) that it will lead me to so act. But then I stop and reflect, recall my dieting plans, and resolve to skip dessert. On the present desire-belief account I had a fleeting intention to have a chocolate dessert. But I am inclined to say that I had no such intention, for I was never appropriately settled in favour of such a dessert" (1987: 20).

Again, intuitions about this example could go different ways. It could be argued that if the agent initially judged in a spirit of resignation that she will eat the chocolate because

[4] See Bratman (1987: 36) for discussion of the notion of 'flat-out belief'. See also Williamson (2000: 99).

of her predominant desire to do so then this sounds rather like she had settled in favour of this option, if only momentarily. However, she just so happened to quickly re-evaluate and re-consider her intention, deciding to skip dessert instead. Indeed, given the fact that what caused the agent to resolve to skip dessert was her recollection of her dieting plans, it might be claimed that it sounds very much as if she realized that the current course of action that she had settled on conflicts with a prior commitment to diet, and decided to resolve this conflict in her system of intentions by adhering to her diet plan. So it seems to me that neither of Bratman's examples are conclusive.

Turning now to Bratman's final objection, the claim that the intention to perform an action is identical with the predominant desire to perform that action, along with the belief that one will perform that action because one predominantly desires to perform it, entails what, for convenience, I will label the 'strong belief thesis':

SBT: For any agent S, S intends to ϕ only if S believes that (I will ϕ).

Bratman's final objection to the supplemented predominant desire analysis is that SBT is false. He argues that in many cases an agent might intend to do something, but be agnostic about whether she will actually do that thing. Bratman considers two kinds of scenarios in which an agent intends to do something, but appears agnostic about whether she will actually do that thing (1987: 37-39). In the first instance, he says that a person might be agnostic about whether she will do what she intends because she is unsure she will succeed in her endeavour. For example, he says that a person might intend to move the log blocking her driveway but be unsure whether she will manage to move the log. Secondly, he argues that a person might intend to do something whilst being agnostic about whether she will even try to do that thing. His example is of someone who intends to stop off at the bookstore but is unsure whether she will because she believes that there is a significant possibility that she will forget[5]. In this second example, since the person is unsure whether she will even remember to stop off at the bookstore it might be doubted whether she does really intend to stop off at the bookstore. However, we might suppose the reason that she is unsure whether she will remember is that she

[5] These arguments can be traced back to Donald Davidson (1980: 50) who argued that often people are unsure whether they will do what they intend because they regard what they intend to do as something difficult. Davidson wrote, "in writing heavily on this page I may be intending to produce ten carbon copies…[but] I do not know, or believe with any confidence, that I am succeeding".

suffers from some sort of mild memory deficit that causes her often to forget such things. In this case, it seems like she really might intend to stop at the bookstore even though she is worried she will forget. There are numerous other examples of this sort that we might think of. Thus we might imagine a group of prisoners who have drawn up a plan of escape. However, they know that the odds are against them and that the likelihood is that all will not go to plan. Do they intend to escape? Yes. Do they believe they will? They are unsure. Indeed, we might suppose that this is their umpteenth attempt this year. So they might have very good evidence that they will fail.

One way of responding to these sorts of examples is to attempt to re-describe the content of the agent's intention in a way that is compatible with thinking that intending to do something necessarily involves believing one will do it. For example, in Bratman's first example, it might be argued that the individual does not intend to move the log *simpliciter*, but intends to *try* to move the log. Since it is plausible that she does believe she will try to move the log, the example, so described, is consistent with SBT.

While perhaps an example such as this, in which a person seemingly intends to do something but is unsure whether she will succeed because she views the task as a difficult one, could be re-described in a manner that is consistent with SBT by appealing to the language of trying, not all Bratman-style counterexamples to SBT are cases in which it is plausible to suppose that the agent believes that she will try. For instance, Bratman's second example is deliberately designed to eschew this type of response. In Bratman's second example a person intends to do something that she is unsure she will even remember to do. In this example it appears that the agent believes that she may not even try to stop off at the bookstore. She believes she may not remember to at all. Thus it is of no help to the defender of SBT to re-describe her intention as an intention to try to stop off at the bookstore. Nor would this sound like a very natural way of expressing what she intends to do. It is not as if the agent is concerned that the task of stopping off at the bookstore is a difficult or challenging one and thus that she may not succeed. She is concerned that she may forget.

Richard Holton (2008) has presented another sort of case in which it might appear that the agent intends to do something that she is unsure she will try to do. These are cases in which an agent purportedly intends to do something that she is unsure she will do due

to weakness of will. For example, Holton imagines a person who intends to return some books to the library, but is unsure whether she will return them because she believes that she might be "seduced by the lovely glint of gold tooling on the books' smooth, cloth-covered boards" into not giving them up (2008: 30). However, in this example it seems open to the defender of SBT to simply deny that the agent really does intend to return the books. After all, if she believes that she may decide not to give the books up because she is so fond of them then it is not clear that she has actually come to a decision to return them in the first place. The defender of SBT could attempt to reinforce this intuition by appeal to Bratman's claims about the functional roles of intention. If the agent is unsure that she will return the books because she is concerned that she will be seduced by their lovely gold tooling then it seems likely that her attitude towards returning the books is one that could not support the various planning-related activities that Bratman emphasizes. At the very least, it looks like she will be unlikely to reason towards further intentions and treat the idea of returning the books as a stable point within a system of larger plans.

Here is a different sort of case in which an agent appears to intend to do something that she is unsure she will even try to do:

S knows that she will be interrogated tomorrow. She knows that certain interrogation techniques will be employed which she will find difficult to resist. Nonetheless, she forms the intention to try to withhold what she knows. However, S also knows that her interrogators have injected her with a truth serum. She knows that in fifty percent of cases the truth serum is effective. The individual injected with it comes to the conclusion that he or she wants to disclose everything. However, in fifty percent of cases the serum does nothing at all. S intends to try not to give up any information when she is interrogated. However, given her statistical knowledge of the effectiveness of the serum, she is agnostic about whether she will try.

Still, even if this example and Bratman's second example are indeed cases in which the agent is unsure whether she will try to perform the action in question, the defender of SBT could argue instead that in these examples the content of the agent's intention is conditional in some way. Thus she might argue that in Bratman's second example, the agent's intention is of the form, 'I will stop off at the bookstore, if I remember'. In the example above, she might argue that the agent's intention is of the form, 'I will try to withhold the information, conditional on the serum not preventing me'. Plausibly, these conditionals are propositions that the agent would believe. So the defender of SBT

might argue that in some purported counterexamples the agent's intention is best characterized in terms of the intention to try, while in others it is better characterized as the intention to ϕ if the relevant conditions are met, or if one can, or if circumstances are allowing, and so on[6].

What seems paradoxical about these responses to the purported counterexamples to SBT is that while we might well expect the agent to say, for instance, "I intend to try to move the log", or, "I intend to stop off at the bookstore, if I remember", in describing what her intention is, either in report to herself or to others, from a third-personal perspective, it also seems correct to say of the agent that her intention is to move the log. So it may seem true to say both that the agent intends to ϕ, *simpliciter*, and that she intends to try to ϕ, or intends to ϕ conditionally. Which of these descriptions of the agent's intention is appropriate seems to be to some extent context-dependent. How can we explain this? I think that those who reject SBT, such as Bratman and Holton, might explain this by arguing that whenever a person intends to do something it is normally just trivially true that she intends to try to do it, or intends to do it so long as everything goes to plan or circumstances are allowing. They might then argue that in some contexts, to say that one intends to try to ϕ, or that one intends to ϕ if one can, or if relevant conditions obtain, will be trivially true, but misleading, while in others it might be informative, because it carries the conversational implication that one thinks that doing what one intends will be difficult and one is not confident of success. On this analysis, saying, for instance, '*S* does not intend to *try* to stop off at the bookstore, she intends to stop off at the bookstore', would be an instance of what Laurence Horn (1989) refers to as metalinguistic negation. Metalinguistic negation occurs whether the utterance of a negative sentence is not intended as a negation of the proposition expressed by that sentence, but rather as a rejection of the way it is being expressed because, though true, the sentence uttered is an understatement or is inaccurate in some way. On this analysis, saying, 'She doesn't intend to try to ϕ, she intends to ϕ', would be a bit like saying, 'She's not good, she's great!', or 'She didn't manage to move the log, she did it easily'. It's not that the agent literally does not intend to try to ϕ. Rather, while the agent might well say she intends to try in order to convey her lack of confidence, the point is that she does

[6] For a philosopher who defends this strategy of re-describing the agent's intention in terms of trying, or as conditional or qualified in some way, in cases in which it might appear that the agent is agnostic or unsure about whether she will do what she intends to do, see Velleman (1989: 113-121).

not *merely* intend to try, but to actually do it. A person *could* merely intend to try to do something (e.g. move the log blocking her driveway) – say, if she did not actually want to do it, but just wanted to demonstrate to someone else how difficult it is. However, the cases we are considering are not like this. The agent does intend to perform the action. She is just unsure she will succeed.

David Velleman (1989), who is a philosopher who defends the strong belief thesis, and whose views I will discuss in depth in chapter three, has a different explanation of why in some cases it seems appropriate to say both that an agent intends to try to ϕ, or intends to ϕ conditionally, and that she intends to ϕ, *simpliciter*. He argues that sometimes we use the terms 'intention' or 'intend' to refer to an agent's 'goal'. He says that an agent's goal is what she is ultimately motivated to do. It is some achievement which she acts 'with the intention' of bringing about. On the other hand, he says that sometimes we use the terms 'intention' and 'intend' to refer to whatever the agent has decided upon doing at the close of deliberation. Velleman glosses this by saying that sometimes we use the terms to refer to what he calls the agent's 'goal-states', whereas at other times we use them to refer to what he calls the agent's 'plan-states' (1989: 112). In the purported counterexamples to SBT, Velleman says that in saying that the agent intends to ϕ (e.g. "move the log"), we are referring to the agent's goal-state. And in saying that the agent intends to try to ϕ (e.g. "try to move the log"), or intends to ϕ conditionally, we are referring to the agent's plan-state. Velleman says that in developing his theory of intention he is concerned with offering an analysis of the latter sense of intention. That is, he concerned with the sense of 'intention' that refers to what course of action an agent has decided upon doing. Velleman rejects the assumption that a theory of intention should offer a unified analysis of the different senses of 'intention'. He argues that the concept of 'intention' is fundamentally polysemous and is simply used to refer to different things. Following Velleman, the proponent of the supplemented predominant desire theory might also argue that she is offering an analysis of an agent's intention in the sense of her 'plan-state'. Appealing to the distinction between plan-states and goal-state, she might attempt to explain why in some cases it can seem true to say both that the agent intends to ϕ *simpliciter*, and that she intend to try to ϕ, or intend to ϕ conditionally, in the same way.

In a review of Velleman's book, *Practical Reflection*, Bratman says of Velleman's distinction between intentions as plan-states and intentions a goal-states that "When we divorce intentions from goals in this way we are in danger of changing the subject" (1991: 125). In fact, when discussing what he calls "the problem of the package deal" (1987: 143) - the problem of how it can be correct to say that an agent did not intend the consequence of an intended action when that consequence was foreseen by her - Bratman himself draws a distinction between what an agent decides or chooses, or what he calls the 'conclusions of her practical reasoning', and what an agent intends. Bratman argues that choices or conclusions of practical reasoning are holistic in nature. As he puts it, "They are just conclusions that a certain overall scenario is superior to its competitors" (1987: 153/4). In other words, he thinks that a person's choices or practical conclusions are *preferences* for total states of affairs. On the other hand, Bratman argues that what a person intends need not be holistic in this way. Rather, according to Bratman, what a person intends is restricted to what she views as ends and means to her ends. Therefore, the content of an agent's intention may be narrower than the content of the conclusion of practical reason from which it issued. It seems to me that Bratman's distinction between choices or conclusions of practical reasoning and intentions is exactly parallel to Velleman's distinction between plan-states and goal-states. Only for Bratman it is the latter, not the former, which is the principal object of philosophical enquiry about intention. However, I think that in claiming that in divorcing intentions from goals Velleman is changing the subject Bratman's prior theoretical commitments are creeping in. For Bratman, we ought to try to understand the nature of intention in the context of our being planning agents. Thus the paradigm of intention is the directing of one's mind, and one's actions, towards the achievement of some future goal. Bratman believes that once we understand this paradigmatic future-directed case we can then attempt to develop a unified theory of the different senses of intention and extend the analysis to the cases of intentional action and acting with an intention. But, as will been seen in chapter three, Velleman's theory begins from a very different starting place. Velleman is primarily concerned with explaining what makes an action intentional rather than what is involved in intending to do something in the future. He just does not accept Bratman's idea that the future-directed understanding of intention as being directed towards some goal is paradigmatic. Exactly the same is true with respect to the methodological commitments of the proponent of the supplemented predominant desire theory. She begins with a certain picture of intentional action and then attempts to extend that

analysis to the case of intending to act. So I think that if Bratman wants to show that the distinction between 'plan-states' and 'goal-states' is in some sense artificial then he would need to show that this whole approach to understanding intention is misconceived.

In conclusion, I think that Bratman's objections to the supplemented predominant desire analysis are inconclusive. Nonetheless, I believe that we already have other grounds for rejecting that analysis. This is because it is not clear how the addition of the belief condition helps to overcome the two supplementary objections raised to the identification of intention with predominant desire at the end of section 1.3. If we understand an intention to act as a predominant desire to act plus a belief that one will so act because of that predominant desire, we will still not be able to explain our ability to form intentions in Buridan situations. In Buridan cases the agent does not predominantly desire any of the options. Yet in such situations we do not normally remain paralyzed. We just decide to do one thing or the other. Nor does it appear that the added belief condition will enable the supplemented predominant desire theorist to explain the norms pertaining to the rational (non-)reconsideration of future-directed intention. Suppose my predominant desire at *t1* to finish the paper I am writing simply gives way to a predominant desire at *t2* to go out for dinner. Further, suppose that at *t1* I believed that I would finish the paper because I predominantly desired to finish it, but at *t2* I formed the belief that I will go out for dinner instead because at this time this is what I predominantly desired to do. There is nothing irrational about my desiring most to finish my paper at *t1* and desiring most to go out for dinner at *t2*. What about my beliefs? My beliefs at *t1* and *t2* are based on my knowledge of what I predominantly desire at each of these times. What I predominantly desire changed across these times. My beliefs simply tracked the evidence concerning what I would do by tracking the change in my predominant desires. However, if I intended to finish my paper at *t1*, but at *t2* reconsidered this intention and formed the intention to go out for dinner instead then, unless there was some relevant change in the evidence or the information, this would seem to involve some kind of irrationality. In changing my mind I am either judging that my earlier decision was mistaken, in which case it seems rational for me to reconsider my intention, or I am not judging that my earlier decision was mistaken, in which case I am open to rational criticism for undermining what by my own lights was a good decision. The supplemented predominant desire analysis seems unable to explain this.

1.4 Intention and Planning

As we have now seen, Bratman's argument against desire-belief reductionism rests on the idea that future-directed intention has certain distinctive marks or characteristics, which all relate to what he describes as its commitment-characteristic. On the one hand, Bratman describes certain *descriptive* features of future-directed intention. These include the dispositions to pursue or act in conformity with one's future-directed intentions, to treat the matter of what one intends to do as settled, to reason towards and intend means to one's intended ends, and to refrain from forming further intentions that one believes are inconsistent with one's prior intentions. On the other hand, Bratman describes certain *normative* features of future-directed intention. He says that there are 'internal' norms governing the 'plan-constrained' perspective of the deliberating agent. These are means-end coherence and the consistency constraints. He also says that there are norms 'external' to the plan-constrained perspective of the deliberative agent, which concern such matters as when to re-open deliberation and what make for good habits of (non)reconsideration of prior intentions.

Supposing that future-directed intention does indeed have these features, this still leaves something unexplained. For the question arises of *why* future-directed intention has these features. Bratman's answer to this further question is that we are planning agents. According to Bratman, mature human beings are creatures that make plans. Planning agency is essential to our survival and form of life. Plans are complex and coordinated representations of our practical futures. Bratman argues that since plans are stable but revocable, and have a partial, hierarchical structure (i.e. they tend to start off incomplete and have a nested structure), and since future-directed intentions are the atoms or building blocks of larger plans, this explains why future-directed intentions have the descriptive properties that he identifies. Bratman also argues that forming plans has certain long-term benefits. It facilitates interpersonal and intrapersonal coordination. It enables our prior deliberations to shape our later conduct. And the partial, hierarchical nature of plans, so the fact that plans start out incomplete and tend to be filled in as we go along, enables us to coordinate our affairs in a way that allows for the unpredictability and contingency of the future (1987: 29/30). According to Bratman, plans will be most conducive to producing these sorts of long-term benefits if they are sufficiently stable to

supports coordination, but tend to be reconsidered under the correct types of circumstances; if they are internally consistent and realizable in the world as the agent believes it to be; and if they are filled in by the agent to a degree of detail sufficient for their successful execution. This is Bratman's explanation of the normative features that he identifies with future-directed intention. In both cases, the direction of explanation is from the nature of plans to the nature of future-directed intention. This is because on Bratman's view plans are 'intentions write large'. According to Bratman, future-directed intentions are plan-states.

I agree with Bratman that the distinctive characteristics of future-directed intention cannot be accounted for if we understand intention either as mere 'predominant desire', or as the combination of a predominant desire to act plus the belief that one will act because one predominantly desires to so act. However, though I believe that intentions are not reducible to any other, supposedly more basic kind of mental state or attitude, or combination of states or attitudes, I think we should question the strong, metaphysical connection between future-directed intention and planning that Bratman posits. My reasons for this are two-fold. The first is that, while it is undoubtedly true that the capacity to form future-directed intentions facilitates our capacity to form plans, it does not logically follow from this that intentions are essentially or metaphysically bound up with planning. Thus we need not follow Bratman in supposing that to be a creature that forms future-directed intention *just is* to be a planning agent. It seems possible to at least conceive of beings who form single, isolated intentions that do not fit into larger plans. Perhaps infants or young children acquire the ability to form intentions for the future before developing the relevant capacities associated with genuine planning agency. And perhaps there are non-human animals that can appropriately be described as intenders, or as having future-directed intentions, but not as planning agents[7]. One way to illustrate this point is through an analogy with episodic memory. We have episodic memories of past events. Often these memories agglomerate in a coherent narrative structure, though often there will also be inconsistencies and gaps in certain parts of the narrative. Perhaps there are also certain normative demands on episodic memory for beings like us relating to consistency and coherence, which derive their normative force from the fact that forming coherent and coordinated autobiographical narratives is beneficial to us in

[7] As David Charles (1989: 49) asks, "Could animals have intentions of a simple type without yet being able to plan? Could this, for example, be true of children at some stage of their development?"

various ways. However, we need not conclude from this that episodic memory is constitutively or essentially narrative in nature, and that a creature can only have episodic memory if it has the capacity to form autobiographical narratives, just as we need not conclude that intention is constitutively or essential bound up with planning.

My second reason for questioning the strong, metaphysical connection between future-directed intention and planning posited by Bratman is that on such a view it becomes difficult to explain the fact that we also form present-directed intentions, or intentions to perform an action now. I have two principal reasons for claiming that we form not only future-directed intentions, but also present-directed intentions. The first relates to the thesis, to be defended in chapter five, that all intentional action involves or is in some way related to some intention *to act*. However, it is evidently the case that not all intentional actions are preceded by some intention to act in the future, for not everything we do is decided in advance of the time of action. In which case, it must be the case that some intentional actions are a manifestation of intentions that are not for the future. Rather, they manifest present-directed intention. The second reason for claiming that we form present-directed intentions relates to Buridan cases. As I have already suggested, one important role played by intention is in resolving situations in which we find ourselves faced with a choice where we see no particular reason for favouring one option over the others. However, such Buridan situations arise not only with respect to choices about the more distant future, but also with respect to choices about how to act in the immediate future. Thus Buridan situations are just as much a reason for positing the existence of a distinctive state of present-directed intention as they are for positing a distinctive state of future-directed intention[8]. However, if intentions are constitutively tied up with planning then it looks like we cannot account for the formation of present-directed intentions. Bratman identifies future-directed intentions with certain normative and descriptive characteristics associated with cross-temporal co-ordination and planning which enable an agent to resolve practical problems in advance of the time of action. However, the norms and regularities in terms of which Bratman identifies future-directed intention have no application in the case of present-directed intention, where the time of the agent's formation of her intention is synchronous with the time of action. Thus it is not clear that there is any room for the formation of present-directed intention on

[8] This point is also made by Holton (2009: 13).

Bratman's account[9]. On Bratman's account, it may be that future-directed intentions transform into present-directed intentions, in the sense that the content of a future-directed intention is temporally updated as time passes and the time of action arrives. However, on the planning theory there seems to be no rationale for forming present-directed intentions that do not issue from some prior future-directed intention.

Bratman appeals to the notion of planning in order to explain why intention has certain features or characteristics. As just mentioned, he argues that forming plans has certain important benefits for creatures such as ourselves. And he argues that plans will be most conducive to producing these sorts of benefits if they are suitably stable and if they are consistent and means-end coherent. Might there not be another way of explaining these features or characteristics of intention that does not appeal to planning agency? Many philosophers claim that belief has an aim. Some argue that the aim of belief is truth; others argue that it is knowledge. I think that it is plausible to suppose that intention also has an aim. The aim or purpose of intention is not truth or knowledge, but self-control. The point of intending to do something is to determine what one will do in the future. If this is right, it seems to me that this would explain how intentions, even isolated or atomic intentions that do not knit together into larger plans, are subject to norms relating to stability, consistency and coherence. If the aim or point of intending to do something is to determine what one will do in the future then this aim is likely to be frustrated or undermined if one's intentions are insufficiently stable and too readily reconsidered, or if they are internally inconsistent or means-end coherent[10]. So this would account for the normative aspects of intending. As Bratman also points out, intentions have a causal role as well. They are conduct-controlling pro-attitudes. They contribute causally to what we do. In the final chapter I will discuss the question of how intentions are causally relevant to action.

1.5 Conclusion

In conclusion, though I reject the strong, metaphysical tie between intention and planning posited by Bratman, I believe that Bratman is correct in thinking that intention is non-analysable either in terms of predominant desire or the combination of the

[9] This objection is raised by Velleman (1991: 278)
[10] The claim that the aim or point of intention is self-control and that this explains the various normative characteristics of intention is also defended by Charles (1989).

predominant desire to act and the belief that one will so act because one predominantly desires to. However, at this point it is important to emphasise that there are a number of other reductionist proposals defended in the literature that Bratman does not consider in his (1987). For example, some philosophers, such as Gilbert Harman (1976) and Kieran Setiya (2007), argue in different ways that an intention to ϕ *just is* a special kind of belief that one will ϕ. Another proposal defended in the literature is that intending to ϕ constitutively involves predominantly desiring to ϕ because one believes that one will ϕ, rather than vice versa. As we will see in chapter three, this is a rough approximation of the account defended by David Velleman (1989). What all these views have in common is that they posit a constitutive relation between intending to act and believing that one will so act. Let us refer to all such views as 'cognitivist'. The term 'cognitivism' was first introduced to refer to a certain type of view about the nature of intention by Bratman (1991). Bratman associates the term with two different ideas: first, that intending to act necessarily involves believing that one will so act. Second, that the rational demands or requirements that intentions be consistent and means-end coherent are derived from corresponding rational requirements on involved beliefs. Chapter two will be concerned with evaluating the latter idea. Generally, throughout the following chapters, I will be using the term 'cognitivist' in a less restricted way so as to refer to anyone who holds the first of these ideas - that intending to act is constitutively tied to believing that one will so act. Though I have argued that we have good grounds for rejecting the view that intending to act involves the predominant desire to act plus the belief that one will so act because one predominantly desires to, this is merely one type of cognitivist theory. A full defence of the irreducibility of intention will have to provide reasons, not just for rejecting this view, but for rejecting all these other cognitivist proposals as well. There are two principal motivations for the theories defended by Harman, Setiya and Velleman. One motivation is the idea that positing a necessary connection between intending to act and believing that one will act can explain the rational demands of intention-consistency and means-end coherence by way of the involvement of belief in intention. However, for Velleman and Setiya, the main motivation for their theories is to explain certain ideas concerning the epistemic conditions on intentional action, which are inspired by Elizabeth Anscombe's (1957) influential, through controversial, thesis that it is the mark of intentional action that when a person acts intentionally she knows what she is doing 'without observation'. In chapters two and three I look at each of these ideas in turn. In

chapter four I offer objections of a general nature to the idea that intending to perform an action constitutively involves the belief that one will perform that action.

Chapter Two

Cognitivism and the Requirements of Practical Rationality

2. Introduction

Most theorists of intention agree that intention is subject to certain rational requirements or demands. The main requirements discussed in the literature are for intentions to be consistent and to be means-end coherent. According to Gilbert Harman (1976, 1986), Kieran Setiya (2007b) and David Velleman (2007), the rational requirements for consistent and means-end coherent intentions are grounded, by way of the involvement of belief in intention, in rational requirements for consistent and coherent beliefs. Bratman (1991) refers to philosophers who try to explain the rational demands for consistency and coherence of intention by appealing to corresponding demands on purportedly involved beliefs, 'cognitivists'. In his paper, 'Intention, Belief, Practical, Theoretical' (2009), Bratman identifies certain problems with the cognitivist explanation of the rational demands for consistency and means-end coherence of intentions. He argues that we cannot explain the demands for intention-consistency and means-end coherence only by appealing to corresponding demands on purportedly involved belief. He argues instead that the requirements on intention are better thought of as distinctively practical in nature and rooted in the long-term benefits of planning agency. My aim in this paper is to assess Bratman's objections to the cognitivist derivation of the requirements on intention. I argue that his objections are inconclusive. It is not my intention to suggest that cognitivism is true – I present other objections to cognitivist theories in subsequent chapters - but merely that it is not clear that the problems that Bratman finds with the cognitivist derivation of the requirements of practical rationality are decisive ones. This chapter has four parts. In section 2.1 I will try to get clear about the cognitivist explanation of intention-consistency and means-end coherence. In section 2.2 I will present concerns that Bratman has about the different possible ways that the cognitivist might attempt to explain means-end coherence. In sections 2.3 and 2.4 I will critically discuss Bratman's concerns, focusing particularly on an assumption that Bratman makes that we are fallible about our intentions.

2.1 The Cognitivist Explanation of the Requirements

Most theorists of intention think there are certain rational demands or requirements which intention is subject to. The main requirements on intention discussed in the literature are,

> *Intention-consistency*: For any agent *S*, at a single time *t*, it is rationally required that (If *S* intends to *A*, and *S* believes that doing *A* and doing *B* are incompatible, *S* does not intend to *B*).
>
> *Means-end coherence*: For any agent *S*, at a single time *t*, it is rationally required that (If *S* intends to *E*, and S believes that if *S* will *E* then *S* intends to *M*, *S* intends to *M*).

The demand for means-end coherence is also sometimes referred to as the demand for instrumental rationality. As this principle is stated, *S* is only required to intend *M* if *S* believes that *intending* M is necessary in order to *E*. The reason for this is simple to illustrate using examples of involuntary means. Suppose that I am an insomniac. I believe that I will not be able to sleep tonight. However, given that I believe I will not sleep, I think that I may as well use the time to finish my essay. I also believe that in order to finish my essay it is necessary that I don't sleep. I am clearly not required to intend to not sleep because I intend to finish my essay and believe that this requires that I stay up. After all, I believe that I will not sleep regardless of whether or not I intend to. Consider a further example. Suppose that I intend to live for another five minutes and also believe that this requires that I continue breathing. Am I rationally required to intend to continue breathing? We can also illustrate the point by considering examples of foreseen, unintended side effects. Suppose that I intend to chop some wood and believe that if I will chop some wood then I will blunt my axe. Am I rationally required to intend to blunt my axe? For these sorts of reasons, it seems plausible to conceive of means-end coherence as a requirement concerning the relation between an agent's ends and her beliefs about what it is necessary to intend as a means to her ends. One might object that the claim that *S* might have a belief that intending to *M*, rather than doing *M*, is necessary in order to *E* sounds slightly artificial. Perhaps then we might think of S's

belief about *M* along the lines of the following conjunction of beliefs: (If I am going to *E* then I will have to *M*) and (It is not the case that (*M* is just going to happen by accident, independently of my will)).

An important thing to note about these two constraints on intention, as I have stated them, is that they are *wide-scope* requirements on the combination of an agent's attitudes. The requirements are wide-scope because they do not state that there is any particular one of the attitudes that the agent must have in order to satisfy the requirements. A rational requirement does not entail that the agent has reason to have any *particular* attitude, in the sense of considerations in favour of having some particular attitude that make having that particular attitude in some respect good. Rather, they concern the relation between the collection of an agent's mental states or attitudes. The thought is that if one is in violation of some rational requirement then something has gone wrong with respect to the current *combination* of one's attitudes[1]. There are a few philosophers who have doubted whether there are any such wide-scope, rational requirements. In his (2005) paper, 'The Myth of Instrumental Rationality', Joseph Raz argues that there is no rational requirement for means-end coherence. More recently, Niko Kolodny (2008) has extended Raz's approach to intention-consistency. It is not my aim in this paper to address this issue. Rather, my goal is to assess the arguments presented by Bratman for thinking that *if* there are such wide-scope requirements on intention then cognitivist theories cannot explain them.

If intention is subject to the requirements of intention-consistency and means-end coherence then it seems to be an important question that an adequate theory of intention must address why intention is subject to them. As we have seen, according to Bratman's 'planning theory' of intention, the state or attitude of intention is a distinct member of our catalogue of folk-psychological states, non-reducible to any other state or attitude, or combination of states or attitudes, and subject to its own distinctive and sui generis rational demands or pressures for consistency and means-end coherence. Bratman argues that these rational demands are explained by the fact that intentions are the building blocks of larger plans, and that forming plans has various kinds of important long-term benefits. He thinks that plans will be most conducive to producing these long-term benefits if, among other things, they are internally consistent and realizable in

[1] See, for example, John Broome (1999) and (2005).

the world as the agent believes it to be, and if they are filled in by the agent to a degree of detail sufficient for their successful execution. Bratman's planning theory of intention can be contrasted with what he calls 'cognitivist' theories of intention[2]. Bratman defines 'cognitivism' in terms of the following two ideas,

"The first is that intention seems in some way to involve belief. Different views are possible here: some suppose that an intention to *A* involves the belief that one will *A*; some say that it only involves a belief that *A* is, in an appropriate sense, possible. But however we spell out this belief-involvement we may also be struck by a second idea, namely that one's beliefs are themselves subject to demands for consistency and coherence…And this leads to a conjecture: the rational demands for consistency and coherence of intention are grounded, by way of the involvement of belief in intention, in rational demands for consistency and coherence of belief. To this extent, practical rationality of one's system of intentions is, at bottom, theoretical rationality of one's associated beliefs." (2009: 30)

The first idea that Bratman associates with cognitivism is that intention involves belief. He distinguishes between three versions of this idea that have appeared in the literature,

"(a) intending to act just *is* a special kind of belief that one will so act; (b) intending to act *involves* a belief that one will so act; (c) intending to act involves a belief that it is *possible* that one will so act." (2009: 31)[3]

These three different belief-conditions on intention are in descending order of strength. (a) entails (b). However, (b) does not entail (a). The claim that intending to ϕ involves or somehow entails believing that one will ϕ is consistent with thinking that the state or attitude of intending to ϕ is distinct from and non-reducible to any sort of belief. The weakest condition is (c). Bratman observes that a version of cognitivism appealing solely to (c) is unable to account for the demand for intention-consistency. This is because believing that it is possible than one will do *A* and believing that it possible that one will do *B* is compatible with believing that it is impossible that one will do both *A* and *B*. Therefore, (c), in conjunction with the claim that the rational requirements on intention are grounded purely in the rational requirements on involved belief, permits jointly intending two actions that the agent believes are impossible both to perform. Bratman argues that the only way to make a cognitivist explanation of intention-consistency that appeals to (c) work would be to supplement it with a further principle of intention

[2] Bratman first coined the term 'cognitivism' in his (1991).
[3] Bratman attributes (a) to Gilbert Harman (1976) and to David Velleman (1989). Bratman attributes (b) to Harman (1986). Bratman attributes (c) to Jay Wallace (2001). As will be made clear in the following chapter, I think it is not quite accurate to attribute (a) to Velleman.

agglomeration according to which if one intends to do *A* and intends to do *B* then one must intend to do *A* and *B*. It would then be possible to appeal to (c) to argue that if one has this single, agglomerative intention to do *A* and *B* then one has the single, agglomerative belief that it is possible that one will do *A* and *B*. This would explain why the agent's agglomerative intention is challenged by the belief that doing *A* and doing *B* is impossible. However, Bratman observes that such a principle of intention agglomeration could not be derived from theoretical demands on the purportedly involved beliefs. Therefore, the resulting account would not be a 'pure' form of cognitivism, according to which the rational demands on intentions are derived solely from the corresponding demands on involved belief, but a 'supplemented' version of cognitivism that appeals in part to the demand for consistency on involved beliefs and in part to a non-theoretical rationale.

The objection Bratman raises against (c) seems correct. The remainder of this paper will be restricted to examining Bratman's objections to a theorist of intention that make the following two claims,

i. For any agent *S*, *S* intends to ϕ only if *S* believes that (I will ϕ). *(Strong belief thesis)*
ii. The theoretical, wide-scope requirements that govern and constrain the formation of beliefs wholly explain the practical, wide-scope requirements that govern and constrain the formation of intentions.

From here on, when I speak of 'cognitivism', I mean the conjunction of (i) and (ii). For the rest of this chapter, unless specifically qualified, if I speak of 'cognitivism' it should be clear that this is what I have in mind. According to Bratman, both (i) and (ii) are false. In the previous chapter, I considered objections that Bratman raises to (i). Against (i), Bratman presents purported examples in which a person intends to do something but is unsure whether she will do that thing. I suggested that such examples are inconclusive. In chapter four I will present what I believe are stronger reasons for rejecting (i). In this chapter I focus on Bratman's arguments against (ii). That is, I focus on his rejection of the cognitivist's explanation of the rational requirements governing intention.

The cognitivist explanation of the practical, wide-scope requirements governing intention presupposes (i) - the strong belief thesis. The strong belief thesis is identical to

Bratman's belief-condition (b). According to the strong belief thesis, intending to act necessarily involves or entails believing that one will so act. As was suggested previously in section 1.4, the notion in play here is what is often referred to as 'flat-out', or 'outright belief'. What is it to have an outright belief that p? Timothy Williamson (2000) suggests that we think of outright belief as being an essentially practical attitude. Williamson argues that outright believing something involves willingness to take that proposition for granted, or treat it as certain, in one's practical reasoning (2000: 99). A further thing to note about the state or attitude of outright belief is that there seem to be distinctive rules or requirements for how the formation of outright beliefs should go. One such rule is as follows,

> *The entailment rule*: For any agent *S*, at a single time *t*, it is rationally required that (If *S* believes *P*, and S believes that *P* entails *Q*, then *S* believes *Q*).[4]

It might be argued that there is reason to reject the entailment rule as overly strong. Are we rationally required to believe every trivial consequence of everything that we believe? For example, I believe that every proposition entails its own conjunction. So I believe that *p* entails (*p* & *p*), that (*p* & *p*) entails ((*p* & *p*) & (*p* & *p*)), and so on, ad infinitum. By the entailment rule, it would seem that I am thereby rationally required to believe an infinite series of such entailments. This is obviously implausible. Here are three possible replies to this objection. One would be to say that the entailment rule is an idealisation. It is a rational requirement for someone who is ideally rational and not subject to the same limitations in time and mental capacity as we are[5]. A second response might involve making the qualification that the entailment rule is a requirement on an agent's beliefs, but only *when the relevant question is being attended to by the agent.* So for any agent *S*, at a single time *t*, if *S* is attending to the question whether *Q* then it is rationally required that (If *S* believes *P*, and *S* believes that *P* entails *Q*, then *S* believes *Q*). A third possible response is just to emphasise that beliefs are dispositional states. So if you belief that *P*, and believe that *P* entails *Q* then you are rationally required to be *disposed* to take *Q* for granted in your reasoning and planning whenever the question of whether *Q* is relevant to your reasoning and planning. Either way, I don't think that these sorts of concerns are major problems for the entailment rule.

---

[4] I have followed Scott Sturgeon (2008) in calling this the entailment rule. I have stated the entailment rule as a wide-scope rational requirement. Sturgeon himself does not.
[5] See, for example, Setiya (2007b: 666).

Recall that according to the demand for intention-consistency, it is rationally required that (If *S* intends to *A*, and *S* believes that doing *A* and doing *B* are incompatible, *S* does not intend to *B*). Assuming the strong belief thesis and the entailment rule, here is one way the cognitivist explanation of the demand for intention-consistency might go. Call this (A):

1) At time *t*, S intends to *A*. *(Supposition)*
2) For any agent S, S intends to ϕ only if S believes that (I will ϕ). *(Premise: strong belief thesis)*
3) From (1) and (2), at *t*, S believes that (I will *A*).
4) At *t*, S believes that ((I will *A*) entails (it is not the case that (I will *B*)). *(Supposition: i.e. because S believes that A and B are incompatible))*
5) For any agent S, at a single time *t*, it is rationally required that (If S believes *P*, and S believes that *P* entails *Q*, then S believes *Q*). *(Premise: entailment rule)*
6) Given (5), at *t*, it is rationally required that (If S believes that (I will *A*), and S believes that ((I will *A*) entails (it is not the case that (I will *B*)), S believes that (it is not the case that (I will *B*)). *(Instantiation of entailment rule)*
7) So, at *t*, it is rationally required that S believes that (it is not the case that (I will B)).
8) So, given (2), at *t*, it is rationally required that (S does not intend to B).
9) Therefore, at *t*, it is rationally required that (If S believes that (I will *A*), and S believes that ((I will *A*) entails (it is not the case that (I will *B*)), S does not intend to B). *(Conclusion: intention-consistency)*

What about the cognitivist explanation of means-end coherence? Recall that according to the demand for means-end coherence, it is rationally required that (If *S* intends to *E*, and S believes that if *S* will *E* then *S* intends to *M*, *S* intends to *M*). I can see two possible ways the cognitivist explanation of means-end coherence might go, which I will refer to as (B) and (C). As we will see, both (B) and (C) rely on some further principle in addition to the strong belief thesis and the entailment rule. Beginning with (B),

1) At time *t*, *S* intends to *E*. *(Supposition)*

2) For any agent *S*, *S* intends to ϕ only if *S* believes that (I will ϕ). *(Premise: strong belief thesis)*

3) From (1) and (2), at *t*, *S* believes that (I will *E*).

4) *S* believes that ((I will *E*) entails (I intend to *M*)). *(Supposition: means-end belief)*

5) For any agent *S*, at a single time *t*, it is rationally required that (If *S* believes P, and *S* believes that P entails Q, then S believes Q). *(Premise: entailment rule)*

6) At *t*, it is rationally required that (If *S* believes (I will *E*), and *S* believes that ((I will *E*) entails (I intend to *M*)), then *S* believes that (I intend to *M*). *(Instantiation of entailment rule)*

7) So, at *t*, it is rationally required that (*S* believes that (I intend to *M*)).

Thus far, the argument only gets us to the requirement that the agent *believes* that she intends the means. At this point, the argument might go in either of two directions. Firstly, continuing with (B):

8) For any agent *S*, it is necessarily the case that (if *S* believes that (I intend to ϕ) then *S* intends to ϕ). *(Premise: infallibility principle)*

9) From (8), at *t*, necessarily (if *S* believes that (I intend to *M*) then *S* intends to *M*). *(Instantiation of infallibility principle)*

10) So, from (7) and (9), at *t*, it is rationally required that (*S* intends to *M*).

11) Therefore, at *t*, it is rationally required that (If *S* intends to *E*, and S believes that if *S* will *E* then *S* intends to *M*, *S* intends to *M*). *(Conclusion: means-end coherence)*

The second way that I can see the cognitivist explanation of means-end coherence going, (C), is identical to (B), except that (8) and (9) are replaced by (8*) and (9*):

(8*) It is rationally required that (If S believes that (I intend to ϕ) then S intends to ϕ). *(Premise: Intention-belief consistency).*

(9*) From (8*), at *t*, it is rationally required that (If S believes that (I intend to *M*) then S intends to *M*. *(Instantiation of intention-belief consistency)*

(10) So, from (7) and (9*), at *t*, it is rationally required that (*S* intends to *M*).

(11) Therefore, at *t*, it is rationally required that (If *S* intends to *E*, and S believes that if *S* will *E* then *S* intends to *M*, *S* intends to *M*). *(Conclusion: means-end coherence)*

I have presented one possible explanation of intention-consistency, (A), and two possible explanations of means-end coherence, (B) and (C). (A) appeals to the strong belief thesis and the entailment rule. (B) and (C) appeal to the strong belief thesis, the entailment rule and either of two further principles which I have called the infallibility principle (i.e. (8)) and intention-belief consistency (i.e. (8*)). We are now ready to consider Bratman's objections to the cognitivist explanation of means-end coherence.

2.2 Bratman's Objections

Bratman thinks that the cognitivist account of means-end coherence runs into difficulties for the reason that we can have false beliefs about our own intentions. That is, he argues that the principle I have stated in (8), which I have called the infallibility principle, is false. He writes,

"Suppose that I intend E and know that E requires M and that I intend M. If I still do not intend M my intentions suffer from means-end incoherence. But suppose that, while I in fact do not intend M, I nevertheless falsely believe that I intend M. So my beliefs are that E, that E requires M and that I intend M, and that M. There is no incoherence (though there is falsity) in this structure of beliefs. So means-end coherence is not belief coherence." (2009: 37)

Bratman says that the cognitivist might respond to this problem by appealing to some further rational demand that is violated when one believes that one intends *M* but does not does in fact intend *M*. In other words, the cognitivist might reject (8) in favour of (8*). Bratman attributes this strategy to Jay Wallace (2001). Explaining Wallace's view, Bratman writes,

"Wallace does not say that intention to A requires a belief that one so intend. But he does aver that if 'intentions are readily accessible to consciousness' then it is 'independently irrational for you to have false beliefs about the content of your intentions' (2001: 22). As for 'intentions that are not readily accessible to consciousness', Wallace says that it is 'doubtful that intentions that are cut off in this way from conscious belief really do introduce rational constraints on our further attitudes, of the kind represented by the instrumental principle' (2001: 23)." (2009: 38)[6]

Bratman rejects the appeal to the further rational demand, which I have stated as (8*), to avoid false beliefs about one's intentions. To begin with, he notes that,

"not all cases of false belief are cases of irrationality. So we need to ask why we should say that, in the present case, the false belief does involve irrationality." (2009: 38)

To put his point slightly differently, (8*) is a particular instance of the more general schema,

> *(RB)* For any agent S, it is rationally required that (if S believes *p* then *p*).

However, (RB) is generally false. Therefore, it seems plausible to suppose that (8*) is false as well. What is required then is a reason to think that, though (RB) is generally false, it is not false when '*p*' is replaced by '(S intends to ϕ)'. Therefore, (8*) is an exception to the general falsity of this schema. Bratman writes, "One idea might be that false beliefs about the contents of one's attitudes always involve irrationality" (2009: 38). In other words, it might be argued that (RB) is true whenever the values for '*p*' are filled in by first-personal attributions of mental states or attitudes. Thus it might be argued that (8*) is a specific instance of the more general schema,

> *(RB*)* For any agent S, and any proposition *p*, it is rationally required that (if *S* believes that *S* herself is *F* then *S* is *F*),

where the values for *F* are filled in by first-personal attributions of mental states or attitudes. However, Bratman argues to the effect that (RB*) also looks generally false.

[6] It should be noted that while Wallace appeals to something like (8*) in his explanation of means-end coherence, Wallace himself would not accept (C) because he rejects the strong belief thesis in favour of the weak belief-condition (c). Wallace fails to realize that (c) does not suffice for the purposes of cognitivism because it is unable to account for intention-consistency. For a cognitivist who appeals to both the strong belief thesis and something like (8*) in order to explain the demand for means-end coherence, see Setiya (2007b). Setiya claims, "there is something incoherent about the belief that I intend to ϕ, unless it is constituted by the intention to ϕ. It is an inherently defective belief" (2007b: 671).

Bratman's reasoning here is roughly as follows. It cannot be a rational requirement of belief that if one believes a proposition then it is true. A false belief may be strongly favoured by the evidence, and may be the product of a perfectly good process of reasoning. On the other hand, a true belief may be the product of a bad process of reasoning. The evidence could be overwhelmingly against it. The fact that the belief is true could just be lucky. Bratman argues that the same holds for when the values for '*p*' are filled in by the first-personal attribution of mental states. He claims that we do sometimes have false beliefs about our mental states or attitudes, such as what we want, fear and love. However, he argues that we are psychologically complex beings with limited time, capacities and resources. The mere fact that we are sometimes wrong about these sorts of things does not necessarily make us irrational. But if it does not seem necessarily irrational to be mistaken about the contents of these sorts of attitudes, why should it always be irrational to be mistaken about our intentions?

Bratman argues that since there does not seem to be anything necessarily irrational about having false beliefs about one's own attitudes, it will be difficult to explain why it should always be irrational to have false beliefs about one's intentions only by appealing to the rational requirements on belief. If we want to justify (8*), Bratman says it will be necessary to treat (8*) as a basic requirement of practical rationality that is independent of the rational demands on belief. As a case in point, it is worth noting that Bratman's planning theory of intention would seem to entail the rational requirement stated in (8*). According to the planning theory, intention is a sui generis state or attitude, non-analysable in terms of desire and/or belief, and subject to its own distinctive norms of practical rationality, which include consistency of intention and means-end coherence. With regards to the demand for consistent intentions, recall from chapter one that Bratman distinguishes between the demand for internal consistency and the demand for strong consistency of intention with the agent's beliefs. Since strong-consistency requires that it be possible to successfully execute one's plans given that one's beliefs are true, and since having false beliefs about one's own intentions is likely to scupper one's efforts at the effective realisation of one's plans as much as any other kind of false belief, it would seem to follow that strong-consistency requires consistency between one's intentions and one's beliefs about one's intentions. According to Bratman, these requirements are not grounded in purportedly involved beliefs, but obtain in virtue of the various roles that intention plays in facilitating planning agency. A different argument

for intention-belief consistency, which also does not appeal to the purported involvement of belief in intention, can be constructed appealing to the claim proposed in chapter one that the aim or purpose of intending to do something is self-control. If the aim or purpose of intending to do something is to determine the actions of one's future self then this aim is likely to be frustrated if one has false beliefs about one's intentions. Bratman says that the cognitivist might herself appeal to non-theoretical considerations of some sort in order to justify (8*). In this case, (C) could only deliver what Bratman calls a 'supplemented' version of the cognitivist explanation of means-end coherence. It would have to appeal to both rational requirements of theoretical rationality and independent rational requirements of practical rationality. Such a supplemented cognitivism would endorse (i) above, i.e. the strong belief thesis, but reject (ii), the claim that the theoretical, wide-scope requirements that govern and constrain the formation of beliefs *wholly* explain the practical, wide-scope requirements that govern and constrain the formation of intentions. Rather, the thought would be that the fact that the strong belief thesis is true only *partly* explains the rational requirements on intention. The requirements on intention are *not* purely the consequences of the theoretical requirements on belief. Bratman says that this is still a coherent position, but it will be less appealing for the cognitivist. Presumably, it will be less appealing because if the cognitivist concedes that non-theoretical considerations, such as those relating to the long-term benefits of planning, or to an aim or purpose of intention pertaining to self-control, must do some of the work in explaining the rational requirements on intention, there would seem to be a pressure to allow that such non-theoretical considerations might fully explain the rational requirements on intention. However, if the cognitivist allows this much, simplicity would then seem to favour the view that such non-theoretical considerations are doing all of the explanatory work, and that there is no reason to also appeal to the purported involvement of belief in intention in order to explain these requirements.

2.3. Infallibility

In this section and the next I take a closer look at Bratman's objections to the cognitivist explanation of means-end coherence. I will mainly discuss (B), and specifically Bratman's assumption that we are fallible, so can have false beliefs, about our own intentions. In this section I will argue that, contrary to what Bratman suggests, it is not

at all obvious that (8), the infallibility principle, is false. I suggest that it is perhaps possible to conceive of certain types of cases in which a person has false beliefs about her intentions, though I myself am undecided about what such purported counterexamples to (8) ultimately show. I argue that even if such purported counterexamples to (8) do show that it is possible to have false beliefs about one's intentions, the strangeness of these cases supports the intuition that there would indeed be something irrational about it.

Bratman does not offer any positive argument in favour of the assumption that we are fallible, so can have false beliefs about, our own intentions. Is this an assumption that Bratman is entitled to help himself to without argument? I think that Bratman might argue that whether he owes some sort of positive argument for thinking that we are fallible about our intentions, or whether the onus is on the cognitivist to show that this assumption is false, comes down to a dialectical issue about where the burden of proof lies. Bratman might argue that just as the principle of intention-belief consistency is a particular instance of the more general schema, (RB*), and this more general schema is false, so the infallibility principle is a more specific instance of the following more general principle,

(Q) For any agent *S*, necessarily (if *S* believes that *S* herself if *F* then *S* is *F*).

where the values for *F* are filled in by first-personal attribution of mental states or attitudes. Bratman could argue that (Q) is generally false. What reason is there for thinking that (Q) is generally false? As Timothy Williamson puts it, with respect to pretty much any proposition you can think of, including first-personal attributions of mental states, "Mistakes are always possible. There is no limit to the conclusions into which we can be lured by fallacious reasoning and wishful thinking, charismatic gurus and cheap paperbacks" (2000: 94). The thought then is that it seems conceivable that one can pretty much come to believe anything, including specious claims about the nature of one's own attitudes. For this reason, Bratman could claim that it is not true that a person's beliefs about what she wants, fears, believes, etc., can never be mistaken. Bratman could argue that since (Q) is generally false then, given that (8), the infallibility principle, is a more specific instance of the schema stated in (Q), it is justified to assume that (8) is also false. So he could say that the onus is on the other side to offer some

positive reason for thinking that there is something special about the case of beliefs about one's own intention that makes it an exception to the general rule.

Is there some positive reason for making an exception for intention – for thinking that even if it is not generally true that we are infallible about our own mental states and attitudes, it is true with respect to our intentions? To begin with, it will be helpful to draw a distinction made by Richard Moran (2001: 55-65) between two different ways in which one can seek to acquire knowledge about one's own mind, or to answer the question of, for instance, what one thinks about *x*, or what one feels about *y*. Sometimes we ask such questions in a 'deliberative' spirit. For example, if I ask the deliberative question, 'Do I believe that *p*', my aim is to acquire knowledge about my own beliefs by making up my mind about the matter. Moran expresses this by saying that the deliberative question, 'Do I believe that *p*', is 'transparent' to the question of whether *p* is true (2001: 60). In making up my mind about whether *p* is true I immediately acquire the answer to the question whether I believe that *p*. At other times, we ask questions such as, 'Do I believe that *p*', 'Do I want to ϕ', etc., in what Moran calls a 'theoretical' spirit. In the theoretical spirit, we seek to acquire self-knowledge by interpreting the evidence, in the same manner in which we attempt to answer questions about the attitudes and states of mind of another. In this theoretical spirit, I might even consult a third party, such as a therapist, in the hope that she is able to correctly interpret my first-personal reports on my outward behaviour and inner, mental life and help me to gain better insight into myself. Sometimes the deliverances of this theoretical perspective will directly conflict with the deliverances of the deliberative perspective, showing one to be inconsistent or incoherent in one's attitudes. Furthermore, in the theoretical mode of self-enquiry, there is always the possibility that one will come to a conclusion on the basis of mistaken inference, or on the basis of the mistaken inferences of some other person in whose authority one trusts. It seems to me that, in order to defend (8), the infallibility principle, the cognitivist needs to argue that, unlike one's other states or attitudes, one can only acquire beliefs about what one intends in *one* of these ways – namely, through the deliberative route. This would mean that the *only way* in which one can form the belief that one intends to ϕ is by actually making up one's mind on the matter. In which case, the only way that one can meet a rational demand or requirement to believe that one intends to ϕ is by making up one's mind that this is what one is going to do.

Is it plausible to suppose that an agent can only form beliefs about what she intends deliberatively, by making up her mind on the matter? At the least, I think that it is plausible to suppose that this would be the normal way to acquire belief about what one intends, and that there would be something strange or odd about attempting to answer the question of what one intends in a third-personal, 'theoretical' manner. The reason is that whether a person possesses the intention to ϕ seems to be up to her in a way that whether she has, say, some belief, or some desire, may sometimes not be. To a certain extent we are passive with respect to what we desire, or feel, or believe about some matter. And our passivity with respect to these sorts of states or attitudes makes them potential sources of self-estrangement. They are not things that are always fully under our control, making them the potential objects of self-deception or repression. Hence we may enquire about them in the third-person-like or 'theoretical' spirit identified by Moran. However, I think that there is something very strange about the idea of enquiring about what one intends in this sort of way. For instance, it would seem peculiar to approach one's therapist with the question of what one intends to do, in the way that one might approach one's therapist with the question of what one wants or believes about some matter. It would be very odd to attempt to answer the question, 'Do I intend to ϕ?' in the same sort of way that one might attempt to answer the question, 'Am I addicted to cigarettes?'. What a person intends just seems to be a matter for her to decide. The answer to the question 'Do I intend to ϕ?' seems to depend entirely on, and be transparent to, one's own conscious choice or decision.

The idea that what a person intends to do is up to her in a way that her other attitudes may not be seems to be built into the theories of a number of cognitivists. As will be explained in more detail in the following chapter, Velleman (1989) and Setiya (2007) both argue in different ways that forming an intention to ϕ involves jumping to an as yet unsupported, causally self-fulfilling belief that one will ϕ because one wants or is motivated to ϕ and believes that forming the belief that one will ϕ will dispose one to ϕ. On Velleman's and Setiya's accounts, whether one intends to ϕ is necessarily a matter of one's consciously endorsing one's reasons or motivations for ϕ-ing by jumping to the as yet unsupported self-fulfilling belief that this is what one is going to do, in order to *thereby* do it. Whatever one thinks about the details of this sort of account, on this picture, what one intends to do is purely a matter for one's own conscious decision. Nothing in

Bratman's account commits him to this sort of picture of intention as an attitude of conscious endorsement of one's reasons or motives for acting. Bratman's functionalism about intention entails that whether or not one has some intention fundamentally depends on various modal truths about how an agent would be disposed to reason and act under various types of circumstances. A person might well be mistaken or ignorant of such truths. Thus it is hardly surprising that Bratman thinks that a person could easily be ignorant or mistaken about her intentions.

Bratman might respond to the argument above by conceding that there is something strange about a person approaching the question of what she intends in what Moran calls the theoretical spirit. However, he could argue that the fact that there is something strange about it does not show that it is impossible. Here Bratman might return to Williamson's argument that "mistakes are always possible". Bratman might argue that confronted with her guru, a person might reason as follows:

1. Everything that my guru tells me is true.
2. My guru tells me that I intend to ϕ.
3. Therefore, I intend to ϕ.

To illustrate, imagine a subject who is told by a group of scientists that they have both the technological capability and the know-how to detect patterns in her brain states and directly read off the corresponding mental states, including her intentions, which are realized in those brain states. Further, suppose that this information is in fact false. The scientists lack this ability. Nonetheless, impressed by the scientists' credentials, the subject fully believes what she has been told by them. She has total confidence in their authority with respect to her own mind, including her intentions. Suppose that the scientists tell the subject that she has an intention to ϕ. Bratman could argue that it is at least conceivable that the subject might believe that she intends to ϕ on the basis of what she believes to be the scientist's authority, but that she might not in fact intend to ϕ because ϕ-ing is something that she judges she has no good reason to do, or is even something that would be positively bad to do. Let us suppose that at a later time the subject forms an intention to ω. She knows that if she is going to ω then she intends to ϕ. She believes that she intends to ϕ. However, she does not really intend to ϕ. In

which case, she has means-end coherent beliefs, but she does not have means-end coherent intentions.

The underlying thought that this example is intended to illustrate is that with sufficient trust in some designated authority, a person could come to believe pretty much anything, including false claims about her intentions. In fact, I find the example quite difficult to imagine. Nonetheless, let us suppose that in these sorts of circumstances in might be possible for you to falsely believe that you currently intend to do something. Your guru tells you that you intend to ϕ. You do not intend to ϕ. Nonetheless, you believe that you currently have this intention because your guru says so. Even if the cognitivist concedes this much, the problem for Bratman is that intuitively there does seem to be something very strange about being in this position. Therefore, even if such examples do indicate that it is possible to form beliefs about one's intentions in a third-personal, theoretical spirit, and to be mistaken, they also seem to yield the intuition that there would indeed be something irrational about this. In which case, the cognitivist simply has reason to move from (8), the infallibility principle, and explanation (B) of means-end coherence, to (8*), the requirement of intention-belief consistency, and explanation (C). Recall that Bratman's objection to (C) is that it does not in general seem necessarily irrational to be mistaken about the contents of one's own attitudes. Therefore, in order to justify (8*) a special case would have to be made for the irrationality of having false beliefs about one's intentions. Bratman argues that since there does not seem to be anything necessarily irrational about having false beliefs about one's other attitudes, it will be difficult to explain why it should always be irrational to have false beliefs about one's intentions only by appealing to the rational requirements on belief. If we want to justify (8*), Bratman says it will be necessary to treat (8*) as a basic requirement of practical rationality that is independent of the rational demands on belief. Here the cognitivist might argue that the reason it is necessarily irrational to have false beliefs about one's intentions is simply that the only thing that could possibly count as grounds for the belief that one intends to do something is the fact that one has made up one's mind, or reaffirmed one's previous decision, to do it. Once again, this is because what a person intends is up to her in a way that her other attitudes are often not. It may be perfectly rational to adopt a third-personal, theoretical stance towards the question of what one believes, fears, desires, and so on. However, there does seem to be something inherently strange about adopting such a stance towards the question of what one intends.

2.4 Further Objections to Infallibility

In this section I consider some alternative reasons that might be given for supposing that we are fallible about our own intentions. In recent years what might be characterised as broadly Cartesian ideas about a person's epistemic relation to the contents of her own mind has come under attack from two main directions. One has been arguments brought by Williamson (2000) against what he calls the 'luminosity' of the mental. The other is recognition among psychologists and scientists of the existence of both unconscious mental states and subpersonal cognitive processes. Does either of these broadly anti-Cartesian ideas have any bearing on the question of whether we can have false beliefs about our intentions? I will begin with Williamson's arguments against 'luminosity'. In *Knowledge and its Limits*, Williamson rejects the following principle, (L),

> "(L) For every case α, if in α C obtains, then in α one is in a position to know that C obtains." (2000: 95)

(L) is intended to capture the idea that at least with respect to certain 'core' mental states, such as feeling cold, if one is in the mental state in question then one is in a position to know that one is in that mental state, which is to say that one will know that one is in that mental state so long as one attends to the question of whether one is in it. Williamson thinks that this is false. He argues that for any given mental state, it is conceivable that one might be in that mental state and yet not be in a position to know that one is. The crucial passage of his argument is as follows,

> "Consider a morning on which one feels freezing cold at dawn, very slowly warms up, and feels hot by noon. One changes from feeling cold to not feeling cold, and from being in a position to know that one feels cold to not being in a position to know that one feels cold…Suppose that one's feelings of heat and cold change so slowly during this process that one is not aware of any change in them over one millisecond. Suppose also that throughout the process one thoroughly considers how cold or hot one feels. One's confidence that one feels cold gradually decreases. One's initial answers to the question, 'Do you feel cold?' are firmly positive; then hesitations and qualifications creep in, until one gives neutral answers such as 'it's hard to say'; then one begins to dissent, with gradually decreasing hesitations and qualifications; one's final answers are firmly negative. Let t_0, t_1,..., t_n be a series of times at one millisecond intervals from dawn to noon. Let αi be the case at t_i ($0 \leq i \leq n$). Consider a time t_i between t_0 and t_n, and suppose that at t_i one knows that one feels cold." (2000: 96/7)

Williamson then appeals to the following premise:

"(1i) If in αi one knows that one feels cold, then in $\alpha i+1$ one feels cold." (2000: 97)

Williamson argues that this premise, in conjunction with (L), leads to a false conclusion. To see why he thinks this, consider a moment very early on in the sequence, $t_0...t_n$, when one is clearly in a position to know that one feels cold and, since one is attending to the question, one does know that one feels cold. It follows from (1i) above that at the next moment in the sequence one also feels cold. However, if one also feels cold at this next moment then, given (L), one is a position to know that one feels cold. Since one is considering the question and so knows that one feels cold, once again, it follows from (1i) that at the next moment in the sequence one feels cold too. However, if we keep applying these two premises, it will follow that one feels cold at the next moment in the sequence, and the next, until eventually we arrive at noon when it is demonstrably false that one feels cold. Since this pair of premises leads to a clearly false conclusion, Williamson argues that one of the premises must be false. According to Williamson, the premise to be rejected is (L).

It is not my intention to assess Williamson's argument for anti-luminosity. I simply want to address whether it has any bearing on the question of whether we are fallible, so can have false beliefs, about our own intentions. It seems to me that it does not. Firstly, suppose that Williamson's anti-luminosity argument could be made to work for the state of intention. It simply does not follow from the conclusion that intention is not a luminous state that we are fallible about our intentions. It is perfectly consistent to think both that it is possible to intend to ϕ, but not be in a position know that one intends to ϕ, *and* that it is necessarily the case that if one believes that one intends to ϕ then it is true that one intends to ϕ. Secondly, while Williamson's anti-luminosity argument might work for the majority of mental states such as feeling cold, or being in pain, it seems to me that it has no application to the state of intention anyway. Intending to perform an action is not a matter of degree. One cannot intend to do something more or less, in the way in which a person may feel more or less cold, or be in more or less pain. There is no conceivable temporal sequence in which I start off fully intending to ϕ, progressively intend to ϕ less and less, millisecond by millisecond, until by the end I have ceased to intend to ϕ altogether. Even Bratman would seem to be committed to rejecting the idea of this sort of partial intending. Bratman argues that intending to ϕ involves a kind of

commitment to ϕ-ing, which enables both the agent herself and others to plan, organise their affairs and form further intentions on the basis of the expectation that the agent will do what she intends. A mere partial intention to ϕ could not support this expectation, and therefore could not support the various planning-related activities that Bratman associates with intention[7].

Moving on from Williamson's anti-luminosity considerations, what about the possibility of unconscious intentions? Does this have any bearing on the infallibility principle? Bargh et al. (2001) conducted a series of experiments that might be interpreted as providing evidence for the existence of unconscious intentions. In one of their experiments, participants were presented with a series of scrambled sentences from which they had to extract as many words as possible. Some participants were presented with sentences containing words that primed for performance, such as 'win' and 'strive'. Other participants were presented with sentences containing words that were neutral in this regard, such as 'carpet' and 'shampoo'. Participants in the former category were found to perform significantly better on a subsequent word forming task that they were given, possibly suggesting that they had formed the goal of performing well. Further, subsequent questioning indicated that the participants were not aware of any relationship between the initial word-forming task and their performance in the subsequent one. Bargh takes this to indicate that the behaviour in the primed group was guided by goals that were unconsciously acquired and that the participants were at no point aware of having. If so should this cause us to question the infallibility principle? As Bayne (2013: 172/3) argues, it is not clear whether the behaviour of the participants in the primed group in Bargh's experiment was guided by an unconsciously acquired goal or intention to perform well in the subsequent task, or whether their behaviour was simply modulated by factors of which they were not conscious. Bayne points out that being unaware of the reason why one is motivated to act a certain way is not the same as acting with an unconscious intention. Thus it seems open to interpretation precisely what such experiments show. However, setting this aside, it seems clear that establishing the existence of unconscious intentions still would not show that it is possible to have false beliefs about one's own intentions. It would certainly show that intention is not a

[7] Holton (2008, 2009) has defended the view that there can be 'partial intentions'. However, by an intention that is 'partial' Holton does not mean that the agent is less than fully committed to doing what she intends in a manner analogous to having less than full belief in a proposition, but that the intention constitutes a competing sub-plan for achieving some end if other means of achieving it fail. This is a very different idea.

luminous state. For if one can unconsciously intend to ϕ then one can intend to ϕ without being in a position to know that this is one's intention. It would also show that we are not *omniscient* with respect to our intentions. If a person is omniscient with respect to a proposition *p* then it is necessarily the case that if *p* is true then she believes that *p*. The existence of unconscious intentions would show that it is possible to intend to ϕ, but not believe that one intends to ϕ. However, it would not show that it is possible to believe that one intends to ϕ, but not intend to ϕ. We can illustrate the fact that failures of omniscience with respect to *p* do not entail fallibility with respect to *p* through an analogy with mathematical truths. Since mathematical truths are necessary truths it follows that it is necessarily the case if one believes the mathematical truth that *p* then *p*. Therefore, we are infallible about mathematical truths. However, it clearly does not follow that we are omniscient about mathematical truths.

I have argued that neither Williamson's anti-luminosity considerations, nor the possibility of unconscious intentions, have any bearing on the question of whether we are fallible about our own intentions. I now want to consider whether it might it be possible to reject infallibility about intention in a more direct way, by appeal to more everyday sorts of examples. Unfortunately, Bratman does not offer any concrete examples of mistaken first-personal self-attribution of intention. However, I think it is possible to imagine the sorts of cases that Bratman has in mind. What I will now argue is that such purported counterexamples to the infallibility principle can be at least as plausibly described in a way that is compatible with that principle. Therefore, they do not show that we can have false beliefs about our intentions.

Generally, I think that purported counterexamples to (8) will be of the following kind: Suppose that I have to make a decision about what to do this evening. Either I will stay in and finish my paper or I will go out to a party. I know I cannot do both. If I will go out then I will need to shower, and I will also need to make sure I have my invitation. On the other hand, if I will stay in and work on my essay then I will need to clear my desk and sit down to work. Suppose that I believe that I intend to stay in and finish my paper. However, instead of clearing my desk and getting on with it, I decide that it will be a good thing to take a shower first. Next, suppose that after my shower I go to clear my desk so that I can sit down to work. Amongst the clutter, I discover my invitation for the party this evening. If I intend to stay in and work, I have no need of the

invitation and can throw it away. But instead of throwing it away I place it safely in the draw. I have now done two things which are both essential for going out to the party, but are non-essential to, perhaps even obstructive of, what I believe is my plan to stay in and finish my essay. What is the explanation? One possible explanation is that I am deceiving myself about my intention to stay in and finish my essay. My belief that I intend to stay in is false. However, an alternative, and I think at least equally plausible, explanation is that I do intend to stay in and finish my essay, but I am not resolved on doing this.

Consider a second example. Every Monday morning I see a poster on my way to the library. The poster asks passers-by to make a charitable donation to some cause. When I see the poster I think that it would be good to make a donation. But I am very busy in the week. So I form the intention to make the donation the following weekend when I have less to do. However, when the weekend arrives I never act on this intention. Instead, I always come up with some excuse for delaying making my donation until another time. What is the explanation of my behaviour? One possible explanation is that I am deceiving myself about my intention to give to charity. I believe that I intend to make a donation, when really I do not. However, an alternative, and I think at least equally plausible explanation, is that I do really form the intention to give to charity, but I am not resolved on doing this. Every week I form the intention to make a donation the following weekend, but I never carry my intention out.

My description of these examples presupposes a distinction between intention and resolution. This distinction has also been made by Bratman (1998) and Holton (2009). What reason is there for supposing there is such a distinction? I think that the distinction is natural and intuitive. Very often we form intentions to do things without thinking about the need to try and stick to them. On other occasions, we know that we will be tempted to do other things, but believe that it is important that we stay the course. In such cases, we resolve to stick to our intention. I suggest that we think of resolution as a form of what Bratman refers to as 'personal policy' (1987: 56). According to Bratman, a personal policy is an intention to form a further intention when one finds oneself in particular kinds of circumstance. Bratman gives the example of the personal policy of turning down a second drink when one has to drive home (1987: 57), and a further example would be the policy of putting on one's seatbelt when one gets in a car. I suggest that resolutions are personal policies to resist the temptation to reconsider an

intention. When one resolves to ϕ one forms a policy-intention to resist reconsidering one's intention to ϕ in circumstances in which one feels tempted to reconsider. As already mentioned in the previous chapter, in order to stick to intentions, sometimes we use the kinds of methods of pre-commitment, or self-binding, highlighted by Elster (1979). This can involve acting in ways that makes it practically difficult if not impossible to pursue a course of action, thereby narrowing one's future options. Or it may involve acting in a way that alters the incentives. However, while we do often use these kinds of methods of pre-commitment in order to help us maintain our resolve, on other occasions, it seems clear that we do not resort to such methods. Rather, we aim to stick to our intention simply through resolve[8].

Returning to the examples, my suggestion is that one way of describing them would involve the distinction between intention and resolution. In the first example, it is not that I falsely believe that I intend to stay in and finish my essay. I do intend to do this, but I am not resolved about doing it. I am tempted to reconsider my current plan and go to the party instead. This leads me to act in a way that keeps my options open. In the second example, it is not that I falsely believe that I intend to make a donation the following weekend. I really do have this intention each week. But I never follow through. Again, I am irresolute. Each time I end up putting it off. So these purported counterexamples to (8) strike me as inconclusive. They are not clearly inconsistent with the infallibility principle.

[8] It is a further question how we manage to maintain our resolutions in the face of contrary desires and inclinations. According to Holton (2009), maintaining resolve in the face of contrary desires and inclinations requires utilizing one's 'faculty of willpower'. Holton claims that willpower is a 'faculty' in the sense that it is an ability or capacity that is dedicated to resisting temptation and blocking reconsideration of intentions in general. Holton argues that willpower operates like a 'muscle'. He says it is "something that it takes effort to employ, that tires in the short run, but that can be built up in the long run" (2009: 120). He maintains that if an individual has strong willpower her chances of resisting the temptation to reconsider her intention will be increased, whereas if she has weak willpower her chances of resisting the temptation to reconsider her intention will be decreased (209: 113). Holton appeals to the empirical literature on the phenomenon of 'ego-depletion' in defense of this claim. The research on ego depletion shows that activities requiring self-control draw on finite resources that can become temporarily depleted. Further, when these mental resources are depleted, engaging in further activities requiring self-control becomes more difficult. Holton takes studies that show that ego depleted subjects are less likely to maintain prior resolutions as evidence of a faculty of willpower that, like a muscle, can be temporarily weakened by its exercise. For critical discussion of Holton's notion of a faculty of willpower and his use of psychological research on ego-depletion, see Levy (2011).

2.5 Conclusion

Bratman rejects cognitivism on the grounds that the rational requirements for consistent and coherent beliefs could not wholly explain rational requirements for consistent and means-end coherent intentions via the involvement of belief. In particular, Bratman argues that cognitivism will not be able to explain the requirement for means-end coherence. His argument turns on the assumption that we are fallible about our intentions. In this chapter I have questioned whether this is an assumption that Bratman can help himself to without further argument. I have argued that it is not. The idea of forming the belief that one intends to do something without forming or reaffirming the intention to do it does indeed seem very peculiar. If it is indeed possible in certain rare sorts of cases then there certainly does at least seem something inherently strange or irrational about it. I have suggested that it might be thought that we can reject infallibility on the grounds of broadly anti-Cartesian ideas concerning Williamson-style anti-luminosity considerations or the possibility of unconscious intentions. However, such ideas have no bearing on the issue of infallibility of belief about intention. Furthermore, I have argued that purported examples of mistaken first-personal self-attribution of intention of a more everyday kind can be plausibly re-described in a manner that is compatible with the infallibility principle by appealing to the distinction between intention and resolution. I conclude that if we are to reject cognitivism about intention then it will have to be for other reasons.

Chapter 3

Congitivism and the Epistemic Conditions on Intentional Action

3. Introduction

In this chapter I present objections to the theories of two prominent cognitivists, David Velleman (1989) and Kieran Setiya (2007). In developing their theories, Velleman and Setiya take as their starting point certain ideas, inspired by Elizabeth Anscombe (1957), concerning the epistemic conditions on intentional action. In her monograph, *Intention*, Anscombe argued that it is the mark of intentional action that when a person acts intentionally she knows what she is doing 'without observation'. In order to explain what they take to be correct about this idea, both Velleman and Setiya argue that an intention to act must either be identical with or constitutively involve a special sort of belief that one will so act. In this chapter I make two central claims. Firstly, Velleman's and Setiya's strategy for explaining the proposed epistemic conditions on intentional action is problematic. This is because both fail to adequately explain how it could be possible, or for that matter permissible or warranted, to form the beliefs that they claim are involved in forming intentions. Secondly, both Velleman's and Setiya's ideas about intention, and about the knowledge that we have of our intentional actions, involve a considerable departure from Anscombe. Anscombe herself did not subscribe to the thesis that intending to act constitutively involves believing that one will so act. Anscombe thought that the knowledge that we have of our intentional actions is a distinctively practical species of knowledge that is different from the theoretical or speculative species that Velleman and Setiya treat it as. The chapter proceeds as follows. In section 3.1 I present Velleman's theory. In section 3.2 I explain how Velleman attempts to address the problem of how it could be possible to form the beliefs that he claims are involved in intention. In section 3.3 I present an objection to Velleman's response to the problem. In section 3.4 I explain Setiya's theory and his own response to what he sees as the problem, not so much of how it could be possible, but of how it could be warranted to form the beliefs that he, like Velleman, claims are constitutively involved in intention. I argue that Setiya's response is also unsatisfactory. Finally, in section 3.5 I return to Anscombe. I present textual evidence that indicates that Anscombe did not think that intending to act constitutively involves believing that one

will so act. I offer my own interpretation of what Anscombe meant in claiming that we have non-observational knowledge of our intentional actions.

3.1 Velleman's theory

My aim in this section is to present Velleman's theory of intention. However, I will begin by locating his theory within the context of certain Anscombian ideas about the nature of intentional action. At the beginning of *Intention*, Anscombe draws a three-fold distinction between different ways in which the concept of 'intention' is employed (1957: 1). She says that sometimes we give 'expressions of intention', as in, "I am going to do such-and-such", sometimes we speak of action as intentional, and sometimes we speak of the intention with which something was done. After an initial preamble on "the verbal expression of intention" (1957: 5), Anscombe considers the question of what distinguishes actions that are intentional from those that are not (1957: 9). Her answer is that intentional actions are actions "to which a certain sense of the question 'Why?' is given application" (1957: 9). Anscombe writes that the relevant sense of the question 'Why?' is the sense in which the question would be refused application by the answer, 'I was not aware I was doing that' (1957: 11). She then makes the claim that a person's intentional actions are a member of the class of things that that person "*knows without observation*" (1957: 13). Anscombe maintains that a person's actions are intentional under some description, 'ϕ-ing', only if she possesses a certain kind of knowledge that she is ϕ-ing that is 'non-observational'. Anscombe does not offer any a priori reason in support of this claim. Rather, she appeals to reflection on examples, such as the following:

> "Say I go over to the window and open it. Someone who hears me moving calls out: What are you doing making that noise? I reply 'Opening the window'. I have called such a statement knowledge all along; and precisely because in such a case what I say is true – I do open the window; and that means that the window is getting opened by the movement of the body out of whose mouth those words come. But I don't say the words like this: 'Let me see, what are my movements bringing about? Ah yes! The opening of a window'." (1957: 51)

Anscombe writes that a person's intentional actions are not the only things that she knows without observation. She claims that one knows without observation what she calls the 'mental causes' of things that one does – by way of illustration of a mental cause, she gives the example, ""The martial music excites me, this is why I walk up and

down"' (1957: 16). Anscombe says that a person also knows the position of her limbs without observation (1957: 13). For example, she says that a person usually does not need to be shown the position of her leg in order to know what position it is in (1957: 13). Anscombe also writes that a person can know without observation the reflexive or involuntary movements of her body (1957: 15). Thus she says that one can know without observation and with one's eyes shut that one has kicked one's leg when, for instance, one's knee has been tapped by the doctor (1957: 15). Though Anscombe is not explicit about what she means by 'without observation', I suggest that for Anscombe what knowledge of all these things shares in common is that it is not acquired through any form of direct perception of what is known, nor is it inferred from some further, distinct fact or facts known through the operation of some perceptual faculty. This latter condition excluding *inference* from observed facts is suggested by Anscombe's remark that one does not know the position of one's leg by observing a 'sign', such as a tingling in one's knee, and then drawing an inference about what position one's leg is in from this (1957: 13).

Velleman also claims that when one acts intentionally one possesses what he characterizes as a kind of "spontaneous" knowledge of what one is doing (1989: 4). More specifically, Velleman says that when one acts intentionally one knows what one is doing under some description, such as 'Walking home', or 'Taking a stroll' (1989: 16), that indicates, even superficially, something about one's reasons for one's current actions, or what it is one is ultimately trying to do or achieve (1989: 20/21). Velleman argues that in order to acquire this knowledge one does not need to observe one's bodily movements, compile some sort of account of one's outward behaviour, and then, by putting this together with one's introspective knowledge of one's inner thoughts and feelings, infer what it is that one is up to (1989: 18). He says that one just knows that one is, say, walking home, or taking a stroll, 'spontaneously', without needing to look or introspect.

Nonetheless, Velleman claims that there are two respects in which Anscombe's phrase, 'knowledge without observation', is a misleading characterization of the knowledge that we have of our intentional actions. Firstly, Velleman says that it is misleading because in order to know what one is doing "a background of observational knowledge" (1989: 20) is normally required. For example, Velleman says that in order to know that one is

currently opening the window one will require background knowledge about one's environment and one's abilities, such as that there is in fact a window in front of one, that it can be opened in the current circumstances and that one is capable of opening it (1989: 20). However, in claiming that knowledge of what one is doing often requires some sort of background of observational knowledge, Velleman is not actually in disagreement with Anscombe. At one point, Anscombe writes,

"When knowledge or opinion are present concerning what is the case, and what can happen – say Z – if one does certain things, say ABC, then it is possible to have the intention of doing Z in doing ABC; and if the case is one of knowledge or the opinion is correct, then doing or causing Z is an intentional action, and it is not by observation that one is doing Z" (1957: 50).

Here I take it that Anscombe is saying that a background of observational knowledge about one's environment – i.e. "what is the case' – and one's capabilities – i.e. "what can happen" – is usually a precondition of one's forming the intention to act (e.g. the intention to open the window) in the first place, and in this sense is presupposed by one's having knowledge of what one is intentionally doing[1]. Without this sort of knowledge (e.g. there is a window in front of one, one is capable of opening it) one would probably not have decided to perform the action. However, this still does not mean that one needs to observe oneself performing the action in order to know that one is doing it.

The other respect in which Velleman thinks the phrase, 'knowledge without observation', is misleading is that it does not adequately distinguish knowledge of one's own intentional action from other things known non-observationally. For example, with respect to non-observational knowledge of the position of one's own limbs, Velleman claims that one detects or discovers what position one's limbs are in, not necessarily by observing or inferring, but by 'feeling' what position they are in (1989: 22). However, Velleman says that what one is currently trying to do – e.g. walk home - is not something that one detects or discovers. He says that it is not something that "occurs to you at all" (1989: 22). However, in objecting to Anscombe's use of the phrase, 'knowledge without observation', as misleading in this respect, I think that Velleman is overlooking the fact that Anscombe does not introduce the phrase for the purpose of characterizing the knowledge we have specifically of our intentional actions, for which she reserves the

[1] This subtler aspect of Anscombe's view is noted by Moran (2004: 49).

term, 'practical knowledge', but in order to delineate an area for investigation without begging certain questions that she is interested in pursuing. The purpose of her introducing the phrase is to describe a class of actions without invoking the concepts of the intended or the intentional, the voluntary or involuntary, or of acting for reasons – the very concepts that Anscombe wants to elucidate.

The problem or explanatory task that Velleman sees himself as facing is to explain how we can possess spontaneous knowledge of our intentional actions if our actions are merely the effects of the prior motivating causes of our desires and beliefs. As he puts it,

"We usually know what we're doing, and we seem to know it quite spontaneously, without having to discover it. We feel as if we're inventing what we do and hence that we don't have to find it out. But if our actions are caused by our motives, then they must follow a predetermined course that isn't really ours to invent: we aren't the authors of our actions but merely spectators. How, then, can we have spontaneous knowledge of them?" (1989: 4)

Velleman suggests that, if the traditional Humean causal picture of desire/belief motivation is right, we should expect it to be the case that we are first determined to act as a result of antecedent psychological causes and only then come to realize what we are doing in acting, in the manner of spectators of our actions rather than authors. His task as he sees it is to identify what needs to be added to the traditional Humean causal picture in order to account for our experience of ourselves as having spontaneous knowledge of our intentional actions.

Velleman's solution is to posit an intrinsic desire, or 'sub-agential aim', for self-knowledge or self-understanding that he suggests all rational, self-conscious agents possess (1989: 27) – a desire or sub-agential aim to know not just what one is doing (e.g. walking along the street), but why one is doing what one is doing (e.g. to go home). Velleman says that the desire, or sub-agential aim, for self-knowledge or self-understanding is not something that we directly feel or experience ourselves as having, as we do with other desires or ends. This is because, a bit like the desire to avoid pain, it is not something that we directly pursue but, rather, something that guides our selection and pursuit of all our other aims and ends (1989: 37). Velleman argues that given this desire or aim we are naturally inclined only to perform actions that we regard as intelligible (1989: 33/4). That is, we are naturally inclined only to perform actions that

we regard ourselves as having motives for performing. Velleman illustrates and defends this idea using the example of the person walking up Fifth Avenue (1989: 15). He argues that as soon as she realizes that she doesn't know why she is walking up Fifth Avenue the normal response would be for her to stop what she is doing altogether. Velleman's explanation of this is that her desire for self-knowledge causes or motivates her to refrain from continuing with an action that she has suddenly come to realize she doesn't understand (1989: 27). And he suggests that given her desire or aim to know or understand what she is doing, she will not start acting again until either she has remembered why she was walking up Fifth Avenue, in which case she may recommence doing that, or she has landed on some alternative course of action which she views as understandable. Of course, one might object to Velleman's claim that the normal response in the circumstances that he describes would be for the individual to stop. Might she not carry on walking up Fifth Avenue instead and try to work out what she was doing by looking for signs around her? However, Velleman argues that in this case the agent would have landed on some alternative course of action that she views as understandable – namely, carrying on walking with the aim of discovering what it was she was doing (1989: 31).

It seems to me that Velleman's use of this example is not a very compelling argument for the idea that we have an intrinsic desire or sub-agential aim for self-knowledge or self-understanding. Just as the normal response of someone who has forgotten why she is doing what she is doing might be to stop what she is doing and try to remember or figure it out, the normal response of a person who has lost track of where she is might be to stop what she is doing and try to figure it out. This does not show that she has an intrinsic desire or sub-agential aim to know where she is. Rather, she needs to know what her current location is in order to reach her ultimate destination. Equally, a person needs to know why she is doing what she is doing in order to know whether her current actions (e.g. continuing down the street rather than taking a left turn she just passed, or entering the café just ahead) are contributing to her getting or doing something that she wants to get or do. If she has forgotten what the reason for her actions is then, presumably, this is why she would refrain from going on until she has remembered or has established some new purpose in acting.

Velleman argues that given this supposed intrinsic desire or sub-agential aim for self-knowledge or self-understanding, the actions which an agent will consider performing at any given time will be narrowed down to the class of actions which she views as intelligible. So any action which the agent views as unintelligible, as something that she is not motivated to do, will automatically be screened or filtered out as inadmissible. However, the class of actions that an agent views as intelligible at any given time will frequently be greater than one. So the question becomes why the agent ends up performing one action rather than another, from the class of possible actions that she regards as intelligible things to do (1989: 58). Velleman's answer is that in deciding to perform one particular action, rather than another, the agent forms an expectation, or forecast – in short, a belief – about how she will act (1989: 59/60). She will already have been motivated to perform this particular action since she regards it as an intelligible thing for her to do. However, once she forms the belief or expectation that she will do it, the desire or aim to know or understand what she is doing will add extra motivational weight to her doing it since it will motivate her to confirm her expectation and act as she believes she will. This added motivation will tip the motivational balance in favour of that particular course of action. So Velleman argues that forming the expectation, or the belief about what she will do, actually disposes the agent to do what she expects. The belief she forms in making a decision or forming an intention will be a causally self-fulfilling expectation about how she will act.

Velleman argues that the beliefs involved in making a decision, or forming an intention, are not derived from or formed in response to sufficient prior evidence for thinking that one will perform the chosen action. The agent does have *some* evidence for thinking she will perform the action in question, since she is aware that she is motivated to perform it. But since the agent does not yet expect or believe that she will perform it her evidence for thinking she will is incomplete. Of all the actions the agent regards as intelligible things that she might do, in forming the intention to ϕ, she must jump to the conclusion that she will ϕ. However, Velleman argues that once she forms the belief, or jumps to the conclusion, her awareness that she expects to ϕ fully justifies and completes her evidence for that very expectation (1989: 60). Velleman argues that this applies not only to the expectation that one is going to ϕ at some later point, but also to the expectation that one is going to start ϕ-ing right now. He claims that one does not have sufficient evidence for expecting that one will start to ϕ right now until one forms the expectation

that one is going to start right now. Once one forms the expectation, one's awareness of the expectation completes the evidence for it (1989: 60).

If the beliefs that Velleman claims are involved in forming intentions are not derived from or formed in response to sufficient prior evidence, why does Velleman think that we form them in the first place? According to Velleman, forming the expectation that one will perform a particular action is a means to performing it. Velleman thinks that in forming an intention, the agent forms the belief that she will ϕ because she is aware that she wants or is motivated to ϕ and knows that forming the belief that she will ϕ will dispose her to ϕ. She forms the belief that she will ϕ in order to *thereby* make it true. Velleman articulates the content of the belief that he claims we form in forming an intention as follows:

"Because I have these motives for doing this, because they have been primed by this awareness of them, and because those predispositions will hereby be reinforced, I'll do it next" (1989: 87).

Note that while Velleman thinks that this is an accurate formulation of the content of the belief that we form when we make a decision or form an intention to act, he does not claim that the agent ever actually articulates any of this either to herself or to anyone else. He thinks that the above is implicit in the forming of an intention, even if it never actually attains linguistic expression (1989: 88). A further thing to note is that Velleman, in contrast to Setiya, whose views I discuss below, does not think that intentions just are special kinds of beliefs. Velleman identifies intentions with "self-fulfilling expectations that are motivated by a desire for their fulfillment and that represent themselves as such" (1989: 109). Velleman's account entails that intending to perform an action involves predominantly desiring, or desiring most, to perform that action because one has formed a belief of the special sort that he describes.

We are now in a position to see why Velleman thinks that he has solved the problem of how an agent who is caused to act by her desires and beliefs could have spontaneous knowledge of her intentional actions. Velleman argues that when an agent forms an intention, she comes up with the idea of doing something - an expectation of acting. This idea or expectation is 'spontaneous' in that it is not derived from sufficient prior evidence. It is a conclusion that the agent jumps to. Once formed, the expectation that

the agent forms in forming an intention will be justified. The agent's evidence for it is completed. It is a justified, true representation of what she will do. Her idea of, saying, going for a stroll, or opening the window, now amounts to knowledge of what she will do. Furthermore, by forming the expectation that she is going to ϕ *now*, the agent will acquire knowledge that she is going to start ϕ-ing right away. Therefore, when she actually starts ϕ-ing, she will know that this is what she is doing. She will know that she is ϕ-ing, not by observing or inferring that she is ϕ-ing, or noticing or discovering it, but because she came up with the idea to begin with.

In the next two sections I consider a problem that arises for Velleman's theory. However, before explaining the problem, and how Velleman attempts to address it, it will be helpful to draw attention to a contrast between the account of intention offered by Velleman and the proposal put forward in chapter one, and touched on in the previous chapter, regarding the aim of intention. In chapter one, I suggested that we think of intention as a mental state that, like belief, has an aim. While the aim of belief is often argued to be either truth or knowledge, I proposed that we think of the aim of intention as self-control. In other words, the whole point of forming an intention is to control or determine how one will act in the future. This proposal is at odds with the theory of intention developed by Velleman. According to Velleman, in forming an intention we are effectively striving after knowledge – knowledge of how we are going to act. In the following chapter I will argue that there is something fundamentally mistaken about this conception of intention. In my view, the aim of *belief* is knowledge. In forming beliefs we aspire to know. However, this is not the aim of intention. The aim of intention is not to know what one will do in the future, even in the imminent future, but to determine what one will do in the future. This is why intention can neither be identical with, nor constitutively involve, any sort of belief.

3.2 Intending and Deciding to Believe

According to Velleman, forming an intention to ϕ involves forming a belief that one will ϕ that one decides to form because one wants to ϕ and believes that forming the belief that one will ϕ will dispose one to ϕ. Velleman's account therefore presupposes that one can form certain beliefs just because one wants or decides to. However, some

philosophers think that there are decisive reasons for thinking that beliefs are not the sort of attitude that can be formed or acquired at will in this way. The classic argument for the claim that beliefs cannot be formed at will, or that one cannot decide to believe, comes from Bernard Williams:

"If I could acquire a belief at will, I could acquire it whether it was true or not; moreover I would know that I could acquire it whether it was true or not. 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 reality. At the very least, there must be a restriction on what is the case after the event; since I could not then, in full consciousness, regard this as a belief of mine, i.e. something I take to be true, and also know that I acquired it at will." (1973: 148)

It is important to emphasize that Williams is not denying that we have any control over what we believe. We clearly can exercise control over what we believe in various ways. One very obvious way we can exercise control over what we believe is by deciding to undertake an enquiry into the truth or falsity of certain propositions – i.e. by actively searching for, or reflecting on, reasons for or against thinking of something as true. Another kind of instance in which we can exercise control over our beliefs is where the evidence depends on our own actions. "Will I complete the marathon?" If I enter the marathon, and then invest time and resources in training for it, then this may give me sound reasons for believing that I will. A further way in which we can exercise control over what we believe is illustrated by the case of Pascal's Wager. Pascal believed that we cannot know if God exists, but that there are pragmatic, or prudential, reasons for believing in God's existence. He thought that the expected benefits of belief outweigh the expected costs. However, Pascal was also alive to the difficulty presented by the notion of wanting to believe in God's existence because one thinks this is the best bet. Therefore, he recommended that one gradually cultivate belief in God by acting as if one believes, developing new kinds of habits and behavior, and, perhaps also, by attempting to forget, or re-interpret, one's earlier strategic thinking. Williams is not denying the possibility of any of the above. Rather, what Williams is attacking is the idea that we can decide to believe "in full consciousness", "just like that". He is rejecting the idea that we can spontaneously and in full-consciousness form beliefs just because we want to.

In the above passage, Williams offers two reasons why he thinks that we cannot spontaneously and in full-consciousness form a belief at will. Firstly, he stipulates that it

is a constraint on belief-acquisition that one cannot form a belief "irrespective of its truth". The phrase, 'irrespective of its truth', might be taken in different ways. However, as I understand him, Williams means that if one is to acquire the belief that *p* one must believe that one currently has grounds for thinking that *p* is true. In forming the belief one must take oneself to be responding to reasons that one presently has for thinking *p* is true. Williams thinks that this constraint would be flouted if one could form a belief at will. The second reason has to do with the sustaining or maintaining of the belief. Williams argues that even if one could form the belief at will, as soon as one remembered that one formed it at will, and so formed it 'irrespective of its truth', this would automatically undermine one's belief. So Williams argues that even if one could form a belief at will the belief would be an unstable one.

Carl Ginet (2001) has defended the possibility of deciding to believe at will by appealing to examples in which someone is anxious about the truth of a proposition, such as, 'I remembered to lock the front door', believes that the evidence she has in favour of the proposition is inconclusive, but who nevertheless manages to end her anxiety by deciding to believe the proposition anyway. However, it seems to me that these sorts of examples are far from conclusive. We might plausibly describe such examples as cases in which a person simply fends off the anxiety or emotion surrounding her concern. Alternatively, we might say that in such examples the agent decides, not to believe the target proposition, but to 'accept' that proposition, where by 'acceptance' it is meant that the agent acts *as if* she knows the proposition is true for practical purposes, treating it as a premise in her planning and practical reasoning[2]. Either way, it seems far from clear that these are genuine counterexamples to Williams' thesis.

According to Velleman, every time we form an intention to act, we spontaneously form a belief about how we are going to act, knowing that at present we have insufficient grounds for the belief. This is not to say that the belief cannot be defeated by subsequent evidence. But just that the fact that the agent is aware that she has insufficient reason for thinking that the proposition is true is no obstacle to her spontaneously forming it. So Velleman thinks that Williams's thesis that we cannot spontaneously and in full consciousness decide to belief is false. He writes,

[2] See Bratman (1992) for a discussion of the attitude of acceptance and its relation to belief.

"if a person thinks of his beliefs as self-fulfilling, then he will see that aligning them with his wishes won't entail taking them out of alignment with the truth, for the simple reason that the truth will take care of aligning itself with them. He can therefore think of himself both as believing what he likes and as believing only what's true. Of course, the dependence between truth and belief in this case doesn't run in the direction that Williams has in mind. What Williams has in mind, when he says that beliefs formed at will would have to be formed irrespective of their truth, is that they could not be formed *because* they were true. He's right, of course, but he's forgetting that such beliefs could be true because they were formed – in which case, their formation wouldn't be irrespective of their truth, after all. Whether a person thinks that he believes something because it's true or that it will come true because he believes it, his belief can still "purport to represent reality" and can therefore qualify as a full-fledged belief" (1989: 128/9).

Crucial to Velleman's defense of the view that forming an intention to ϕ involves forming a belief that one will ϕ at will is the idea that the agent regards the belief as self-fulfilling. Velleman argues that if one regards the belief as self-fulfilling then once one forms the belief it "can still purport to represent reality". For given that the agent regards the belief as self-fulfilling, once she forms the belief she will then take herself to have grounds for thinking that it is true. This is intended to address Williams' worry about the maintenance of a belief formed at will. However, there is still the issue of how the agent could form the belief in the first place. On this matter, Velleman invokes an ambiguity in the phrase, 'irrespective of its truth'. As I understand him, and as Velleman also understands him, what Williams means by this phrase is that if one is to acquire the belief that p then one must currently take oneself to have grounds for thinking that p is true. One must take oneself to be forming the belief that p in response to grounds that one presently has for thinking it true. However, Velleman rejects this constraint on belief-acquisition. He says that "such beliefs [i.e. as Velleman thinks intentions involve] could be true because they were formed – in which case, their formation wouldn't be irrespective of their truth". According to Velleman, what is important is that the subject believes in advance of forming the belief that at no time will the belief be unjustifiably held, and not whether she takes herself to currently have grounds for it. Velleman thinks that if the subject does believe this in advance then she will not be forming the belief 'irrespective of the truth'. He maintains that this is precisely what occurs when we form intentions. So Velleman takes himself to have a principled explanation of both how the subject could acquire a belief at will in forming an intention, and why the belief involved in her intention would be a stable one.

Is Velleman's claim that we can form beliefs at will so long as we regard them as self-fulfilling a plausible one? In order to illustrate the claim, consider the following analogy. Imagine that I have invented a machine that is attached to my brain and is able to read my thoughts. The machine is attached to a box that it fills with coloured balls after a button is pressed. I have no idea what colour balls will fill the box after I press the button. But if I form a belief that the balls will be a particular colour after I press the button then the machine will decode my belief and fill the box with balls of precisely the colour I believe they will be. Would I be able to form a belief about what colour balls will fill the box after I press the button? As I interpret him, Williams' remarks about the psychological constraints on belief-acquisition imply that though I know that whatever I believe will turn out to be true, I would not be able to form a belief in this way. One can only come to believe something in response to perceiving or taking there to be grounds for currently thinking it true. However, there can be no such grounds in this case. Of course, if I was able to form the belief that the balls will be, say, blue this belief would itself be a reason for thinking that the balls will be blue. But, so it might be argued, I could not form this belief in the first place. I might be able to 'accept' the proposition that the balls will be a particular colour, in the sense of deciding to treat the proposition that they are particular colour as a premise in my practical reasoning. And I could pretend, or act as if, I believe they will be a particular colour. But I would not be able to come to believe that they are that colour in the sense of genuinely affirming a proposition that I understand, whilst simultaneously being fully conscious that I currently have no grounds for affirming it[3]. However, the problem is that it is simply open to Velleman to deny this. Indeed, Velleman might argue that the strangeness of the example will be mitigated once we realize that the one way I could resolve this impasse is by *intending*, and so thereby believing, that the balls will be blue.

Velleman's account of what is involved in forming an intention seems reminiscent of a children's story in which the main character believes in fairies because she wants them to exist and believes that their existence depends on her believing in them, or in which she

[3] Rae Langton (2003) rejects Velleman's explanation on grounds of this nature. Langton compares Velleman's account of intention with examples of leaps of faith described by William James in his *The Will to Believe and Other Essays in Popular Philosophy* (1896). Langton argues that a subject's believing that some belief will be self-fulfilling is not enough to make it possible to acquire it spontaneously and at will because the constraint on belief-acquisition identified by Williams would still not be met. However, it seems to me that this is begging the question against Velleman since whether Williams is right is precisely the point in dispute.

believes that the existence of some imaginary realm is conditional on her belief in its existence, or in which she believes that she can fly because she wants to fly and believes that the belief that she will fly will cause her to fly. Velleman might argue that if the notion of deciding to believe something just in order to thereby make it true is incoherent, or problematic, why do we not find these sorts of stories incoherent or problematic? Why instead do they seem to grip our imaginations? My suggestion is that when we think about such stories, what we tend to imagine is the main character intensely trying to *imagine* the relevant proposition. And I think the moral of such stories tends to not really be about belief at all, but about the power of the imagination. So getting stuck on the question of whether they illustrate the coherence of the idea of deciding to form a belief at will seems pedantic and to miss the point. Nonetheless, I do think that we should concede to Velleman that the idea of deciding to believe something in order to thereby make it true is coherent and not obviously impossible. And it may even be that there are certain adults, perhaps adults suffering from certain schizotypal conditions, who go in for precisely this sort of thinking. However, I will now argue that, even if Velleman is right that we can spontaneously and in full consciousness form beliefs at will if we believe that those beliefs will be causally self-fulfilling, this will not enable Velleman to solve the problem of how it could be possible to form the beliefs that he claims are constitutively involved in forming intentions. This is because, for reasons to be outlined in the following section, it is implausible to claim that forming *intentions* necessarily involves forming beliefs that one regards as causally self-fulfilling.

3.3 Can't Epiphenomenalists Form Intentions?

In this section I present an objection to Velleman's solution to the problem of how it could be possible to form the beliefs that he claims are constitutively involved in forming intentions. Velleman's solution to the problem turns on two ideas: firstly, an agent can form a belief at will on the condition that she regards the belief as causally self-fulfilling, and, secondly, forming intentions involves forming beliefs that the agent regards as causally self-fulfilling. In the previous section I conceded that Velleman could be right about the first of these ideas. However, I think there are two reasons why the second is implausible. The first is that it implies that whether or not one can form intentions depends on one's views about the causal roles of one's belief. To illustrate, imagine a staunch epiphenomenalist who firmly thinks that citing an agent's desires and belief

cannot offer a causal explanation of her actions on the grounds that beliefs and desires, and mental states in general, are not the sort of thing that can cause actions. This person believes that in performing any given intentional act it is not the case that what causes her to do that thing is her belief that she will do it. Suppose that this hypothetical epiphenomenalist forms an intention to do something - say, buy some milk. Suppose she believes that she will buy some milk. However, given her commitment to epiphenomenalism, she does not believe that her belief will cause her to buy some milk. According to Velleman's analysis of intention, since this person does not regard her belief that she will buy some milk as causally self-fulfilling, it cannot be true that she has the intention to buy some milk. But this just seems false.

The second, related problem is that, because Velleman uses the claim that forming an intention involves forming a belief that the agent regards as self-fulfilling to explain knowledge of one's intentional actions, it makes the matter of whether one knows what one is intentionally doing dependent on one's views about the causal roles of one's beliefs. According to Velleman's analysis, when one forms an intention to ϕ one forms a belief that one will ϕ. What makes that belief justified, and so a candidate for knowledge, is one's awareness that one's belief that one will ϕ will cause one to ϕ. Return to our hypothetical epiphenomenalist with the intention to buy some milk. Since she is not aware that her belief that she will buy some milk will cause her to buy some milk, but thinks quite the opposite, her belief that she will buy some milk cannot be justified in terms of that awareness. So it is not a candidate for knowledge. Therefore, it follows from Velleman's account that even if it is true that she will buy some milk, she cannot know that this is what she is going to do. Nor can she know that buying some milk is what she is doing when she actually proceeds to carry out her intention. She might conceive of herself as buying some milk. But, according to Velleman's account, she cannot count as knowing this. This also seems false.

One way that Velleman could respond might be to ask whether our staunch epiphenomenalist can really 'live' or 'inhabit' her epiphenomenalism, in the same way that it might be asked whether eliminativists about the mental can really believe their beliefs, or radical skeptics can really believe their radical skepticism. It might be said that what these people have are merely 'philosophers' beliefs'. They do not truly believe the views that they espouse. Rather, it is merely a stance that they adopt in the context of

abstract philosophical debate. It might be conceded that such philosophers' belief is belief *of sorts* on the grounds that it has a distinctive functional role similar to belief, but presumably restricted to the context of intellectual discussion. However, it does not carry the implications of what is ordinarily meant when it is said that a person believes that *p*. On this basis, Velleman might argue that the epiphenomenalist might *say* that she believes that her beliefs about her prospective actions are causally inefficacious. However, in truth, her epiphenomenalism is not something that she can believe wholeheartedly. However, it is important to bear in mind that the example of the staunch epiphenomenalist was merely a way of bringing out the point that a person's views on the causal role of her beliefs cannot have the implications that Velleman's theory entails. This point applies, not just to epiphenomenalists, but to anyone who rejects Velleman's analysis of intention. So it applies to anyone who rejects the claim that when she forms an intention she forms a belief that she will act that is causally self-fulfilling. Perhaps Velleman might want to attribute 'philosophers' belief' to anyone who rejects his analysis of intention – not lived conviction as the objection supposes. Alternatively, perhaps he might want to say that anyone who rejects his analysis of intention is simply incoherent – she believes that her beliefs about how she will act are causally self-fulfilling and she also believes that they are not. However, when we make claims contradicting a person's first-personal self-attribution of belief there generally needs to be some justification for doing so. The grounds will normally be that there is some conflict between the individual's behavior and what she professes to believe. But not only is the idea that forming intentions involves forming such complex beliefs phenomenologically implausible, Velleman has no basis for claiming authority here. Since Velleman has no genuine, *independent* reason for making such claims about a person who rejects his account it seems like the default position is to treat that individual as an authority on her own mental states.

What motivates Velleman's analysis of intention is what he sees as the problem of explaining how we can possess spontaneous knowledge of our intentional actions. Velleman's solution to the problem involves the claim that forming intentions involves spontaneously forming beliefs. In order to explain both how it could be possible for us to form the beliefs supposedly involved in intentions spontaneously and at will, and how such beliefs could amount to knowledge, Velleman claims that they are causally self-fulfilling beliefs that the agent regards as such. I have argued that the idea that forming

intentions necessarily involves forming beliefs that the agent regards as causally self-fulfilling is implausible. It is an implausible consequence of Velleman's theory that a person's views about the causal roles of her beliefs could have implications for whether or not she can form intentions. It is also implausible that they could make a difference to whether or not she can know what she is doing when she acts intentionally. One's views about the causal roles of one's beliefs just seems irrelevant to both one's forming an intention and one's having knowledge of one's intentional actions. So I conclude that the idea that forming an intention necessarily involves forming a belief that the agent regards as self-fulfilling is false. If Velleman dispensed with the idea that the agent regards the belief purportedly involved in her intention as causally self-fulfilling, and retreated to a weaker view according to which forming intentions involves forming causally self-fulfilling beliefs, but which the agent herself may not regard as such, he might still be able to explain how such beliefs could amount to knowledge of how one is going to act. Of course, this would require further work. However, he might argue that the belief that one forms when one forms an intention is a candidate for knowledge given the reliability of the belief-forming process, and so by appealing to some sort of externalist epistemology. Nonetheless, Velleman would not have an explanation of how it could be possible to form these beliefs in the first place. This is a major problem for his account.

3.4 Setiya's Theory

Kieran Setiya, another cognitivist, has offered a different response to what he sees as the problem, not so much of how it could be possible, but of how it could be epistemically warranted to form the beliefs that he associates with intention. In this section I consider Setiya's theory and his response to the worry.

Like Velleman, Setiya maintains that forming an intention to ϕ constitutively involves - indeed, Setiya claims is identical with - forming a causally self-fulfilling belief that one will ϕ that one regards as such. However, there are three major differences with Velleman. To begin with, Setiya does not accept Velleman's idea that the beliefs purportedly involved in intentions motivate action by engaging some desire or sub-agential aim for self-knowledge or self-understanding. Setiya argues instead that intentions are beliefs that, like desires, are intrinsically motivating. On the one hand, he says that an intention

to act "represents its content as being true". This is supposed to make intention a sort of belief. On the other, he says that an intention "has the power to cause or motivate the action it depicts" (2007: 40). This is supposed to make intentions like desires. Setiya concludes, 'Intention is a matter of self-referential desire-like belief (2007: 49). I am not clear why Setiya characterizes intentions as desire-like beliefs, rather than belief-like desires. Or why he does not say that intentions are neither beliefs nor desires, but distinct states that, like beliefs, have a world-to-mind direction of fit, but also like desires, have a mind-to-world direction of fit. This not something that I am aware Setiya actually comments on. However, since my main concern in this section is Setiya's response to the problem of how these causally self-fulfilling beliefs get formed in the first place, and not the precise details of his theory, I will set the issue aside.

A second difference with Velleman is that, as far as I can see, the idea that the agent regards the belief that she forms in forming an intention as self-fulfilling does not play as integral a role in Setiya's account as it does in Velleman's. In fact, Setiya does not offer any argument for it. He simply mentions that other authors have defended it, such as Velleman (2007: 49)[4]. Setiya does not appeal to the idea in order to address the problem of how we form the beliefs that he contends are identical with intentions. As noted, Setiya, in contrast to Velleman, frames this problem, not explicitly as one about how it could be psychologically possible to form the beliefs in question, but as one about how it could be warranted to form them (2008: 397). Setiya writes, "if anything is not epistemically licensed, it is believing something you wish were true…without having evidence to show it is" (2009: 397). In other words, forming a belief, not because you take yourself to have sufficient grounds for think the proposition true, but just because you want it to be true, would be unwarranted. There would be something wrong or objectionable about it. As Setiya sees it, the worry is that forming intentions ends up looking like this kind of objectionable wishful believing. However, Setiya explicitly rejects Velleman's claim that believing that a belief will be self-fulfilling is sufficient to warrant forming that belief. He asserts, "the forming of beliefs without prior evidence is

[4] Setiya also mentions Harman (1976). According to Harman, intentions are causally self-fulfilling beliefs that the agent regards as such and that are the conclusions of practical reasoning. Harman argues that in forming an intention to ϕ the agent concludes her reasoning about what to do by forming the causally self-fulfilling belief that she will ϕ, which she forms in order to thereby make it true that she will ϕ. Though he does not explicitly discuss the problem of deciding to believe, Harman appears to think that what makes it possible and permissible to form this belief is the belief that it will be self-fulfilling. Velleman cites Harman as a main source of inspiration for his theory.

epistemically suspect" (2008: 400). As we will see, Setiya has a different story to tell about what warrants the beliefs that he identifies with intentions.

As with Velleman's theory, Setiya's account of intention is largely motivated by certain ideas about the epistemic conditions on intentional action. However, a third difference between Velleman and Setiya is that, according to Setiya, the epistemic conditions on intentional action are weaker than Velleman suggests. While Setiya thinks that it is often the case that we know what we are doing when we act intentionally, he argues that a condition on intentional action requiring *knowledge* of what one is doing is too strong. In his (2008), Setiya asks us to imagine a person who has been given a paralytic in her arm and believes with irrational optimism that the paralytic has worn off, and for this reason that she has the ability to clench her fist if she decides to do so (2008: 389/90). Setiya says that, as it turns out, the paralytic has in fact worn off. Thus the person's belief that she is able to clench her fist if she decides to do so happens to be true. However, she does not know that she has this ability since her belief in her ability to clench her fist is unjustified. Setiya argues that if such a person were to decide to clench her fist then she would do so intentionally. Further, he says that her intentionally clenching her fist would be accompanied by her belief that she is clenching it. However, he argues that, given that she is irrationally optimistic that her arm is no longer paralyzed, her belief that she is clenching her fist would not amount to knowledge. Setiya also notes another complication with the idea that a person ϕ's intentionally only if she knows that she is ϕ-ing. This is that there appear to be examples in which a person ϕ's intentionally but is agnostic about whether she is in fact ϕ-ing. Setiya considers Davidson's (1980: 50) example of a person who intentionally makes ten carbon copies as she writes, but is unsure whether she is managing to do it. In Davidson's example, the agent is intentionally making ten carbon copies, but does not even believe that she is making ten carbon copies, let alone know that she is (2007: 25; 2008: 390). In light of these sorts of counterexamples to the view that a person ϕ's intentionally only if she knows that she is ϕ-ing, Setiya proposes a weaker condition on intentional action, which in his (2007) discussion he labels 'Belief':

"If A is doing ϕ intentionally, A believes that he is doing it, or else he is doing ϕ by doing other things, in which he does believe." (2008: 390)

Following Setiya's (2007) discussion, let us refer to this principle as Belief. The idea behind Belief is that a person's 'basic' intentional actions – what a person doesn't do anything else in order to do - must be accompanied by the belief that she is performing those actions, even if her more complex, non-basic actions resulting from the basic things that she does may not be. Thus, in Davidson's example, the carbon-copier may not believe that she is making ten carbon copies. However, Setiya thinks that if she is doing this intentionally then she will intentionally be doing certain other things, such as pressing hard on the paper with her pencil, that she believes she is doing, and by means of which she is making ten carbon copies. Contrasting his view with Anscombe, Setiya summarizes his position as follows: "setting aside the claim to knowledge as well as the claim to knowledge without observation – I can only do intentionally what I *think* I am doing" (2007: 25). More exactly, his position is that I can only be doing something intentionally if there is something that I am intentionally doing in doing it that I believe that I am doing. As we will see in the following section, Anscombe would agree with the claim that I can only do intentionally what I think I am doing. However, she would deny that what I think I am doing is a matter of belief about what I am bringing about.

Just as the problem or explanatory task that Velleman sees himself as facing is to explain how we can possess spontaneous knowledge of our intentional actions, the problem that Setiya sets himself is to explain the weaker epistemic condition on intentional action stated in Belief. Setiya anticipates the following two explanations of Belief, both of which he rejects before introducing his own explanation. To begin with, he says that it might be supposed that an agent's belief about what she is doing when she acts intentionally is a conclusion of some sort of inference from proprioceptive knowledge of her bodily movement (2008: 393). Setiya objects to this explanation of Belief using an example taken from Anscombe (1957: 52) involving a person who learns to use a contraption by means of which she is able to keep a balancing object level by lowering a lever at the same rate at which her arm would naturally drop. Setiya argues that in a case like this "my knowledge of what I am doing cannot be derived from a prior belief about the movement of my body" (2008: 394). I am not clear how exactly this example is supposed to refute the explanation of Belief appealing to proprioception. Setiya claims that in Anscombe's example, "I have no premise available (through bodily awareness) from which I could responsibly infer that I am keeping the object level" (2008: 394). But assuming some background knowledge of my circumstances, along with knowledge of

my intention to manipulate the contraption, presumably I could know through proprioception or bodily awareness that I am grasping the lever, and lowering it at the speed at which I drop my arm – all basic intentional actions by means of which I am keeping the object level.

According to the second possible explanation of Belief that Setiya considers, the agent's belief about what she is doing when she acts intentionally is inferred from knowledge of what she intends, together with belief in her ability to do what she intends, where Setiya understands belief in one's ability to ϕ as the conditional belief that one will ϕ if one decides to (2008: 394). On this account, if I am, say, intentionally pumping water into a cistern[5] then I will believe that I am because I can infer it from my knowledge that this is what I intend to do, along with belief in my ability to pump water into the cistern[6]. However, Setiya rejects this explanation of Belief on the grounds that it fails to explain why Belief constitutes a necessary truth. He says that the inference from knowledge of one's intention to ϕ, along with belief in one's ability to ϕ, to the belief that one is actually ϕ-ing might not take place because "I might simply fail to put two and two together" (2008: 394). I find this objection to the proposed account puzzling. Surely if a person knows at a time *t* that she intends to pump water into the cistern at *t*, and believes at *t* that she is able to pump water into the cistern in the sense that she will if she decides to, then, presuming that she is attending to the question of what she is doing at *t*, she is going to believe at *t* that she is pumping water into the cistern. It is not entirely clear to me how the agent could "fail to put two and two together".

Nevertheless, Setiya's own view is that if Belief constitutes a necessary condition on intentional action then this must be because, firstly, the state of intention is necessarily present or involved in intentional action, and, secondly, because intention constitutively involves, indeed is identical with, motivating or desire-like belief about one's actions - either about what one is going to do, as in the case of prospective intention, or about what one is currently doing, as in the case of intention in action. (2008: 396). With respect to his example of the person injected with the paralytic who is irrationally optimistic that she is able to clench her fist, Setiya's analysis is that she believes that she is clenching her fist because she intends to be clenching it and this intention consists in a

[5] This example comes from Anscombe (1957: 37)
[6] For a recent defense of this type of explanation of how we usually have knowledge of what we are doing, or at least of our basic actions, when we act intentionally, see Paul (2009b).

kind of motivating, and so self-fulfilling, belief that she is clenching it. With respect to Davidson's example of the carbon copier, Setiya's thinks, somewhat counter-intuitively, that the carbon copier does not intend to make ten carbon copies since she does not believe that this is what she is doing. Nonetheless, Setiya argues that her making ten carbon copies still counts as intentional because there is something else she intends and so believes she is doing, namely pressing hard on the paper, which she is doing *with the end* of making ten carbon copies (2007: 53/4).

It is worth pointing out that while, in his (2007), Setiya treats Belief as a principal motivation for his cognitivism about intention, in his later (2008) Setiya actually argues that this principle is itself open to counterexamples. Setiya considers a variation on his example described above of the person given a paralytic. In this altered version of the example, the person is cautiously, but not irrationally, optimistic that the paralytic has worn off, and therefore that she is able to clench her fist if she decides to. However, once again, it so happens that the paralytic has indeed worn off. The person decides to clench her fist and does so intentionally. However, given that, in addition to having doubts about her ability to clench her fist, she cannot see or feel her fist, her intentionally clenching her fist is not accompanied by the belief that she is clenching it. Further, Setiya says that in this example it is not plausible to suppose that there is some further, more basic intentional action that she is performing in clenching her fist, which she does believe she is doing. Curiously, Setiya does not seem to view this as particularly problematic for his account of intention. He says that he will assume Belief "For simplicities sake" (2008: 392) in framing his discussion of how it can be justified to decide to form the beliefs that he contends are necessarily involved in intention, but that strictly speaking Belief is false and that we need to weaken the epistemic condition on intending still further:

> "If A is doing ϕ intentionally, A believes that he is doing it or is more confident of this than he would otherwise be, or else he is doing ϕ by doing other things for which that conditions holds" (2008: 391)

As far as I can see, this concession would have serious consequences for Setiya's theory. There are two ways in which Setiya could attempt to explain this extremely minimal epistemic condition on intending. One would be to identify an intention to ϕ, not with the desire-like belief that one will ϕ because one believes that one will, but with some

intrinsically motivating, and so self-fulfilling, desire-like state of increased confidence that one will ϕ. However, even if some such proposal could be made to work, it is not plausible. For as Sarah K. Paul (2009a: 550/1) has argued, if the intention is to be identified with the corresponding degree of partial belief then the intention itself must be equally partial or fine-grained. But as was argued in the previous chapter, intentions generally do not seem to be partial in the way that belief is often argued to be. Further, even if intentions could be partial, it is implausible to suppose the degree to which one intends to ϕ would always correspond to the degree of confidence one has that one will ϕ. The fact that the person in the example is less than fully confident that she will clench her fist does not entail that she is less than fully committed to this course of action. Acknowledging this fact, in a reply to Paul's paper, Setiya (2009a: 130) suggests instead that, rather than supposing that the degree of confidence involved in one's intention corresponds to one's degree of commitment to the intended course of action, we should conceive of intentions as "motivating states that involve at least partial belief". However, if this is Setiya's view then it completely undermines his argument for cognitivism. This is because almost any theorist of intention is likely to concede that when one intends to ϕ one will be more confident that one will ϕ than if one did not so intend. Far from motivating a cognitivist analysis of intention, this is perfectly consistent even with Bratman's planning theory. Thus it seems to me that, as unattractive as Setiya may find it, the only real option available to him would be to defend Belief against his own counterexample by arguing that there is in fact some further, more basic intentional action that the agent performs in clenching her fist, which she does believe she is doing. This will have to be some inner act of volition, or willing, or *trying* to clench her fist.

I turn now to what Setiya thinks warrants the beliefs with which he identifies intentions. Setiya's argument appeals to the notion of 'knowledge how'. He defends the following principle:

"(K) If A is doing ϕ intentionally, then A knows how to ϕ, or else he is doing it by doing other things that he knows how to do" (2008: 404)

By way of illustration, Setiya argues that someone who defuses a bomb may not have known how to defuse it. However, if she defused the bomb intentionally then there must have been something that she intentionally did, say, cutting the red wire, which she

did know how to do. Just as Setiya treats explaining Belief as an adequacy condition on a theory of intention, he also treats explaining (K) as an adequacy condition on a theory of intention (2008: 404).

Setiya's explanation of (K) is that it is the agent's state of knowing how to ϕ, in conjunction with her belief that she is able to ϕ, which, as already mentioned, Setiya understands as the conditional belief that she will ϕ if she decides to, that provides her with the epistemic warrant for forming the intention to ϕ in the absence of sufficient prior evidence for believing that she will ϕ. In making this claim, Setiya takes himself to be simultaneously explaining two puzzles – firstly, how we can be warranted in forming the beliefs that figure in our intentions, and secondly, how to explain the role of knowing how in intentional action (2008: 406).

It seems to me that there are two problems with Setiya's solution to the problem of how it could be warranted to form the beliefs that he claims are identical with intentions. The first is that it entails an implausibly demanding standard of warranted or permissible intention. At one point, Setiya himself asks, "Can't I decide to dance the tango at my wedding…even if I don't know how?" (2008: 406). He asserts that this decision would not be justified. He suggests instead that what I must decide to do is to learn how to dance the tango and to exercise this ability at my wedding, once it is acquired (2008: 406). However, surely I would only decide to learn the tango and exercise this newly acquired ability at my wedding in the first place if I had already decided and made it my goal to dance the tango at my wedding, something that I do not yet know how to do. It is not at all clear why one should accept that there is something wrong with making this my intention simply because I do not yet know how to do it, as Setiya asserts. The second worry with Setiya's proposal is raised by Paul (2009a). Paul points out that at any given moment there may be many actions open to an agent, all of which she knows how to do and believes she is able to do (2009: 556). However, a person's knowing how and believing that she is able to perform any one particular action could provide her with no more warrant for forming the belief that she will perform that action than it does any of the other things that it is equally possible for her to do. It is not at all clear how knowledge that ϕ-ing is *one thing* that it is possible one will do, since one knows how to do it and is able to do it, could license the belief that it is *the thing* that one is going to do.

In the previous section I argued that Velleman does not adequately address the problem of how it could be possible to form the beliefs that he claims are involved in forming intentions. The conclusion of this section is that Setiya's solution to what he sees as the problem, not so much of how it could be possible, but of how it could be epistemically permissible or warranted to form the beliefs with which he identifies intentions is also not satisfactory. As we have seen, Velleman's and Setiya's accounts of intention are motivated by certain ideas, inspired by Anscombe, concerning the epistemic conditions on intentional action. In the following section I consider what Anscombe meant by the claim that we possess a non-observational kind of knowledge of our intentional actions.

3.5 Anscombe on Practical Knowledge

Anscombe maintains that it is the mark of intentional action that when a person acts intentionally she knows what she is doing without observation. Since knowledge is standardly thought of as entailing belief that stands in some appropriate relation to the truth, a natural idea is that if a person's statement of what she is doing expresses knowledge of what she is doing then that statement must express some belief about what she is bringing about or making happen. This is not Anscombe's view. Anscombe writes,

> "I say to myself 'Now I press Button A – pressing Button B – a thing which certainly can happen…here, to use Theophrastus' expression…the mistake is not one of judgment but of performance. That is, we do *not* say: What you *said* was a mistake, because it was supposed to describe what you did and did not describe it [as would be the case if what was said expressed a belief or judgment], but: What you *did* was a mistake, because it was not in accordance with what you said." (1957: 57)

Anscombe claims that if the agent's statement of what she is doing expressed a judgment or belief about is happening then if she were not in fact doing what she says she is – e.g. if she were pressing Button B, not Button A - then we would say that her statement is mistaken. However, Anscombe argues that this is not what we ordinarily say. Rather, we say that it is her *action* that is mistaken, not what she says – e.g. 'You did the wrong thing, you pressed Button B'. Anscombe not only thinks that a person's statement of what she is currently doing is not an expression of belief about what is happening or occurring, but that statements of practical foreknowledge are also not expressions of judgment or belief about what is going to happen or occur. She writes that "'I am going

for a walk – but shall not go for a walk' is a contradiction of a sort", but that it is not "a head-on contradiction" (1957: 92), as it would be if, 'I am going for a walk' were an expression of belief about what will happen in the future. Anscombe also categorically rejects the thesis that intending to act necessarily involves believing that one will so act. At the very end of *Intention*, Anscombe writes that a person can intend to do something – for example, hold out under torture – that she is positively certain she cannot do (1957: 94). According to Anscombe then, intending to act does not entail believing that one will act, and statements of what one is intentionally doing, or what one is going to do, are not statements expressing judgment or belief about what one is bringing about, or going to bring about.

Anscombe is aware that if what one says one is doing is not an expression of belief or judgment as to what is the case then it can seem mysterious how it can be said to express knowledge. She writes,

"Can it be that there is something that modern philosophy has blankly misunderstood: namely what ancient and medieval philosophers meant by *practical knowledge*? Certainly in modern philosophy we have an incorrigibly contemplative conception of knowledge. Knowledge must be something that is judged as such by being in accordance with the facts. The facts, reality, are prior, and dictate what is to be said, if it is knowledge." (1957: 57).

Evidently, when Anscombe says that when a person acts intentionally she knows what she is doing without observation, she is not speaking of the theoretical or contemplative sort of knowledge that has generally been the focus of discussion in contemporary philosophy. She does not mean that the agent's description of her actions expresses a belief that is reliably formed or justified on the basis of evidence. Rather, she is speaking of a species of knowledge of quite a different nature – what she calls 'practical knowledge'.

If Anscombe does not think that a person's statement of what she is doing is an expression of belief about what she is bringing about or making happen then what does Anscombe think it is an expression of? Anscombe thinks that a statement of what one is doing is an expression of intention, but formulated in the present progressive rather than the future tense – e.g. 'I am opening the window' instead of 'I am going to open the window', or 'I am painting the wall yellow', instead of 'I am going to paint the wall

yellow'. According to Anscombe, in saying, for instance, 'I am opening the window', the agent is at once offering a description of her current actions, and expressing what her goal or intention in acting is - what it is she ultimately intends to achieve or bring about in walking across the room, unlatching the window, and so on. This is why Anscombe says that when one makes a false statement about what one is doing the mistake is in the performance, not in the action. Intention has a mind-to-world direction of fit, rather than the world-to-mind direction of fit of belief. In intending to perform an action we aim to bring the world into accordance with our representation of it, rather than to conform our representation of the world to the way the world is. This is the point of Anscombe's famous example of the man going around town with a shopping list (1957: 57). If what the man buys does not agree with the items on the list then the mistake is with his purchases, not with the list itself. This is because "If he made the list…it was an expression of intention" (1957: 56), unlike the list of the detective that Anscombe imagines following him, which is meant to be an expression of the detective's beliefs.

As we have seen, Anscombe does not think that a person needs to observe what she is bringing about in order to know what she is doing. Anscombe thinks that a person can know what she is doing even when she lacks confidence in her ability and cannot be sure that she is succeeding without looking or checking. Anscombe gives the example of a person writing something with her eyes closed. She claims that even though the person may not be confident that what she says she is writing is in fact getting written because, for instance, her pen could have strayed off the page or run out of ink, her statement that she is writing the sentence in question still expresses practical knowledge of her intentional actions (1959: 53). Similarly, in response to Davidson's carbon copier, I believe Anscombe would argue that even though the carbon copier cannot be confident that she is managing to produce ten carbon copies without checking, she still knows without observation that what she is doing is making ten carbon copies.

The reason that Anscombe thinks that a person does not need to observe what she is bringing about in order to know what she is doing is precisely because the act-descriptions employed in statements of what one is doing are expressions of intention that take the present progressive form. Descriptions of actions taking the present progressive form are descriptions of unfolding, temporally extended processes. Following Comrie (1976), we can distinguish between two different types of actions in

progress. On the one hand, there are 'telic' actions, such as writing a PhD, or walking home, that have a specific end point built into them. On the other hand, there are 'atelic' actions, which have no necessary end point, but could go on indefinitely. Examples of atelic actions include walking, sleeping, running, and so on. Present-progressive act-descriptions referring to telic and to atelic actions warrant different sorts of logical inferences. If a person goes walking, but her walking is interrupted because, say, she is struck by lightening then it will still be correct to say that she walked. On the other hand, if a person is walking home, but gets struck by lightening half way it will not be correct to say that she walked home. This is because she never reached her destination. However, in both cases, even though the action was interrupted, it would still be true that the person *was* walking, or *was* walking home. The fact that the person was prevented from completing the process does not make it any less true that this is what she was doing. Hence Anscombe writes,

"A man can *be doing* something which he nevertheless does not *do*, if it is some process or enterprise which it takes time to complete and of which therefore, if it is cut short at any time, we may say that he *was doing* it, but *did not do* it" (1957: 39).

This explains why Anscombe thinks that a person who is unsure of her ability to succeed in doing what she intends to do can still know that she is doing what she intends without observation. The statement, 'I am making ten carbon copies', or, 'I am writing '*p*', are descriptions of unfolding actions in progress. The agent does not need to have finished in order for it to be true that this is what she is doing. Nor does the process need to be free of mistakes or difficulties. The hand of the person writing with her eyes closed may stray off the page. However, unless she gives up, she will soon realize this and attempt to correct the position of her hand by bringing it back onto the page. Even if the agent experiences difficulties or obstacles, so long as she retains the intention to make ten carbon copies, or to write the sentence in question, it will normally be true that this is what she is doing.

It might appear to follow from the above that Anscombe thinks that a person's knowledge of what she is doing is infallible. So long as an agent sustains her intention to ϕ then 'ϕ-ing' is *guaranteed* to be a true description of whatever she is currently doing. Therefore, she will have non-observational knowledge of what she is doing under that description. The agent may take detours, or encounter obstacles to ϕ-ing, but neither is

sufficient to deprive the agent of practical knowledge. However, this would straightforwardly contradict the following remarks, which Anscombe makes early on in *Intention*:

"I say…that we can *know*…[the position of our limbs without observation] and not merely *can say* it, because there is a possibility of being right or wrong: there is point in speaking of knowledge only where a contrast exists between 'he *knows*' and 'he (merely) *thinks* he knows'. Thus, although there is a similarity between giving the position of one's limbs and giving the place of one's pain, I should wish to say that one ordinarily knows the position of one's limbs, without observation, but not that being able to say where one feels pain is a case of something known" (1957: 14).

Whatever one thinks of these remarks, it seems to me that there is no inconsistency here. This is because while Anscombe says that "the rare exception is for a man's performance in its more immediate descriptions not to be what he supposes" (1957: 87), she does think that a person's conception of what she is doing can be flatly mistaken. This is illustrated by her example of saying, 'Now I press Button A', while pressing Button B. True, Anscombe says that here the mistake is not in what is said, but in the performance. Nonetheless, she clearly thinks that this is a case in which a person thinks she is doing something that she is not in fact doing. This suggests that in Anscombe's view there is some further minimal condition that needs to be met, aside from the agent's intending to do what she says that she is doing, in order for her statement of what she is doing to be upheld. It seems to me that this further condition is that what the agent is currently doing must be something that she believes could at least potentially be a means or a stage in the process or chain of events leading to the completion of the intended action. If in acting as one is currently is – e.g. pressing Button B, rather than Button A – one is unintentionally or without realizing it doing something that one believes is not a means to or stage in the process of completing the intended action, then this will be a case in which one thinks that one is doing something that it is not true one is doing. In illustration of this, Anscombe writes,

"Consider the question 'Why are you going upstairs?' answered by 'To get my camera'…if someone says 'But your camera is in the cellar', and I say 'I know, but I am still going to get it' my saying so becomes mysterious; at least, there is a gap to be filled…But if I say: 'No, I quite agree, there is no way for a person at the top of the stairs to get it' I begin to be unintelligible. In order to make sense of 'I do P with a view to Q', we must see how the future state of affairs Q is supposed to be a possible later stage in proceedings of which the action P is an earlier stage" (1957: 36).

According to Anscombe, the agent's statement that she is getting her camera will be a true description of her actions so long as (i) the agent's intention really is to get the camera, and (ii) the agent believes that going upstairs is or at least might be a part of the process of getting the camera – it is something that she is doing *in order* to get the camera. She could be mistaken about this. For instance, she might believe that to get the camera she needs the key to the cellar and that the key is upstairs when in fact it is in the cellar door. In this case, she would be taking an unnecessary detour in going upstairs. However, while she would in fact be do something that is not contributing to her completing the action of getting the camera, she would be doing something that she believes is contributing to getting it. Once she realizes her mistake she will continue her search elsewhere. As long she does not give up, 'getting the camera' will still be a true description of the action that is presently unfolding. On the other hand, Anscombe thinks that if by going upstairs the agent is doing something that she herself admits could in no way contribute to her getting the camera – "'No, I quite agree, there is no way for a person at the top of the stairs to get it"' - then the agent is simply not doing what she says she is.

As we have seen, Velleman would agree with Anscombe that one knows what one is doing in acting intentionally in virtue of the fact that one intends to do it. However, he thinks that it seems mysterious on her account how the conception of one's actions embodied in one's intention can constitute knowledge. A central part of Velleman's project is to explain how an agent's intention can embody knowledge in the standard sense of the term. He wants to explain how the conception of one's actions embodied in one's intention can be sensitive to and supported by evidence and so can be justifiably held, even though intentions are formed spontaneously, and not in response to prior evidence. He wants to explain how what Anscombe calls practical knowledge is in fact another instance of ordinary theoretical knowledge. However, it seems to me that Velleman, and Setiya too, miss Anscombe's basic point, which is that there is a very ordinary type of knowledge – the knowledge that we have of our intentional actions – that contemporary philosophers have overlooked. Anscombe argues that what a person is intentionally doing depends largely on what she intends to do. Her intention to act is the fundamental truth-maker of a statement of what she is intentionally doing. What an agent intends to do is a matter for her to decide and make up her mind about; it is up to

her. This gives the agent a special, though not infallible, first-personal relation to, or authority about, the question of what she is doing. As I understand her, it is because the agent has this special, though fallible, first-personal authority about the question of what she is intentionally doing that Anscombe thinks that it is appropriate to speak of her as having a kind of knowledge.

3.6 Conclusion

In this chapter I have examined the theories of two prominent cognitivists, David Velleman and Kieran Setiya. I have argued that both Velleman's and Setiya's analyses of intention are problematic. This is because both fail to adequately address the problem of how it could be possible, or warranted, to form the beliefs that they claim are constitutively involved in, or identical with, intention. I have also suggested that what leads Velleman and Setiya to develop their theories in the first place is a failure to properly understand Anscombe's remarks about the knowledge that we have of our intentional actions. According to Velleman and Setiya, we can only make sense of the idea that a person does something intentionally if and only if she knows, or, as Setiya more modestly puts it, thinks she is doing it, if acting intentionally necessarily involves having some sort of belief about what one is doing. This leads them to posit a constitutive relation between intention and belief. However, Anscombe held that a person's statement of what she is doing is not an expression of belief at all, but an expression of intention. A person knows what she is doing in large part because she knows what she intends to do. The knowledge that we have of our intentional actions is not an instance of what she calls theoretical or speculative knowledge, but is a distinct kind of practical knowledge.

Chapter 4

Intention, Belief and the Normative Role of Knowledge

4. Introduction

Over the course of the previous two chapters I have considered two principal motivations for cognitivism. One is the idea that cognitivism can explain the rational demands of intention-consistency and means-end coherence governing intention by appealing to corresponding demands on purportedly involved beliefs. The other is the idea, defended by Velleman (1989) and Setiya (2007), that analyzing intention as identical with or as constitutively involving some sort of causally self-fulfilling belief can explain certain ideas, inspired by Elizabeth Anscombe (1957), concerning the epistemic conditions on intentional action. In chapter two I argued that, contrary to Bratman, it might be possible to explain the rational demands of intention-consistency and means-end coherence governing intention by appealing to corresponding demands on purportedly involved beliefs. Nonetheless, this is not a decisive reason in favour of cognitivism. There are other possible explanations of the requirements on intention – for instance, that the aim or point of forming intentions is self-control. In chapter three I argued that Velleman's and Setiya's strategy for explaining the proposed epistemic conditions on intentional action is problematic. I have not yet argued directly against the very idea of cognitivism.

My aim in this chapter is to outline some objections to cognitivist theories of intention that are both novel and of a more general nature. According to cognitivist theories of intention, intending to perform an action necessarily involves believing that one will perform that action. The objections I will outline below appeal to the idea defended by a number of philosophers that there is a fundamental norm, or as it also sometimes referred to in the literature, a 'standard of correctness', for belief. Traditionally, the standard of correctness for belief has been said to be truth. Call this the 'truth view' (TV). More recently, a number of philosophers have argued that the standard of correctness for belief is knowledge. Call this the 'knowledge view' (KV). I argue that (KV) is irreconcilable with cognitivist theories of intention. Since (KV) is true we should reject cognitivism. In section 4.1 I introduce a distinction between two forms of cognitivism about intention, 'weak cognitivism' and 'strong cognitivism'. In section 4.2 I

discuss (TV) and (KV). In section 4.3 I present positive arguments for (KV) given by Timothy Williamson. In section 4.4 I argue that (KV) is consistent with our ordinary practices of criticizing and evaluating beliefs. In section 4.5 I present objections to both weak cognitivism and strong cognitivism appealing to (KV). In section 4.6 I answer some potential replies.

4.1 Two Types of Cognitivist Theory

According to cognitivist theories of intention, intending to perform an action necessarily involves believing that one will perform that action. We can distinguish broadly between two varieties of cognitivism, 'strong cognitivism' and 'weak cognitivism'[1]. According to strong cognitivist theories, an intention to ϕ is identical with a special kind of belief that one will ϕ. For example, Harman (1979) and Setiya (2007) argue that having an intention to do something consists in having a certain kind of causally self-fulfilling belief that one will do that thing. These philosophers are strong cognitivists. On their view, an intention to act *just is* a kind of belief.

In contrast to strong cognitivists, weak cognitivists maintain that while having an intention to ϕ constitutively involves believing that one will ϕ, intention is not reducible to belief. Weak cognitivism is compatible with a dual-component analysis of intention according to which an intention to act is reducible to the *combination* of some sort of belief that one will act plus some further component psychological state or attitude. According to one variety of weak cognitivist theory, discussed in chapter one, an intention to ϕ consists in the psychological complex of a predominant desire to ϕ plus the belief that one will ϕ because one predominantly desires to. As discussed in the previous chapter, Velleman (1989) maintains that intending to ϕ consists in predominantly desiring to ϕ because one believes that one will and believes that forming this belief will cause or motivate one to ϕ. Alternatively, in his 1971 British Academy Lecture, H.P. Grice argued that an intention to act consists in the psychological complex of the state or attitude of 'willing' to act - where by willing to act a certain way I understand him to mean some sort of motivating judgment or conclusion that acting that way is the best thing to do – plus the belief that one will so act because one wills to.

[1] I am borrowing the distinction between 'strong cognitivism' and 'weak cognitivism' from Sarah K. Paul (2009b: 3).

Weak cognitivism is also compatible with thinking that an intention to act is a distinct and irreducible state or attitude, non-analyzable in terms any other kind of mental state or attitude, or combination of attitudes, but one that is necessarily accompanied by a belief that one will perform the action that one intends to perform. I am not aware of a philosopher who defends this final view. However, there is nothing incoherent about it.

Cognitivism has been rejected by a number of philosophers over the years, including Donald Davidson (1971, 1978) and Michael Bratman (1987). As explained in chapter one, Bratman argues that intention is a distinct and irreducible planning state that is in no way constitutively tied to having the belief that one will act as one intends. Though I reject the strong, metaphysical connection between intention and planning defended by Bratman, I am in agreement with him that intention is not analyzable in terms of any other, supposedly more basic mental states, and that the cognitivist's thesis that intending to act necessarily involves believing that one will so act is false. In this chapter, I outline objections to cognitivist theories of intention in both their strong and weak forms.

4.2 The Fundamental Norm of Belief

According to strong cognitivists, intentions are special kinds of beliefs. If intentions were identical with beliefs, even special kinds of belief, as strong cognitivists suggest, we would expect intentions and beliefs to have the same essential properties or characteristics. Strong cognitivists maintain that there are differences between intentions and ordinary beliefs. It is in virtue of these differences that they claim that intention is a special kind of belief. Nonetheless, even if intentions were beliefs that are different or unique in some way they would still have to share whatever properties are essential to and individuating of belief, otherwise it would simply be incorrect to call intentions any kind of belief. They would simply not belong to that genus of state or attitude. One characteristic that is thought by many philosophers to be essential to something's being a belief is that it is subject to a certain fundamental norm or 'standard of correctness'. Such philosophers would endorse the following essentiality claim:

> *(EC)* For any mental state *M*, necessarily (if *M* is any kind of belief then *M* is subject to the standard of correctness of belief).

Traditionally, it has been argued that the fundamental norm or standard of correctness of belief is truth. Call this the 'truth view' (TV)[2]. More recently, it has been argued that the fundamental norm, or standard of correctness, of belief is not truth, but knowledge. Call this the 'knowledge view' (KV)[3].

In general, claims about the standard of correctness for belief are of the form,

> (SCB) For any agent *S*, and any proposition *p*, it is correct for *S* to believe that *p* if and only if condition *C* obtains,

where 'condition *C* obtains' is a place-holder for '*p* is true', '*p* is known by S', etc.

The notion of the 'correctness' of an agent's having a certain belief, as it is employed in (SCB), may not seem particularly perspicuous. One might be inclined to think that the claim that a belief is correct simply means that it is true. In which case (TV) would be trivial. So it seems like the notion of correctness needs to mean something more substantive than simply, 'is true'. One way of understanding claims about the standard of correctness for beliefs might be in terms of what an agent 'ought' to believe[4]:

> (SCB') For any agent *S*, and any proposition *p*, it ought to be the case that *S* believes *p* if and only if condition *C* obtains.

The problem with (SCB') is that it looks too strong. The sufficiency part of the bi-conditional requires that subjects actively seek to believe propositions for which *C* obtains. In the case of (TV) this would require that subjects actively seek the truth, even extremely trivial truths that have no significance for them. A weaker, more plausible version might strip (SCB') of this sufficiency claim:

---

[2] Contemporary discussion of the idea that the norm of belief is truth goes back to Bernard Williams' influential paper, 'Deciding to believe' (1973).

[3] A leading proponent of this view is Timothy Williamson (2000), especially ch.1 and ch.11.

[4] In the discussion to follow, I explicate the concept of 'correctness' in deontic terms. In so doing, I am following a general trend in the literature. One might think that the concept of correctness should not be analyzed in terms of more basic deontic concepts, but should be treated as a primitive, evaluative notion. If this is right it does not make a substantive difference to the argument to follow. The objections to cognitivist theories that I present in section 4.5 below could be reformulated using the schema stated in (SCB) without loss of argumentative force.

(SCB") For any agent *S*, and any proposition *p*, it ought to be the case that *S* believes *p* only if condition *C* obtains.

Alternatively, we might hold onto the bi-conditional, but cast the principle in terms of 'permissibility' – i.e. in terms of what an agent 'may' or is 'allowed' to believe:

(SCB*) For any agent *S*, and any proposition *p*, it is permissible for *S* to believe *p* if and only if condition *C* obtains.

Claims about what it is permissible for a subject to believe should be distinguished from claims about what it is reasonable for a subject to believe. Whether a given condition obtains, be it truth, knowledge, or anything else, may not be transparent to the subject. Given the general possibility of failures of luminosity, we want to say that the subject's believing that *p* can be reasonable, even if in some sense it is ultimately wrong. The word 'reasonable' could be exchanged with 'excusable' without loss of meaning. The intuitive thought here is that the agent is not open to blame or criticism. An agent might violate the norm, but their violation of the norm might be entirely in good faith.

Taking (SCB*) to be an adequate schema for interpreting claims about the standard of correctness for belief, the truth view entails,

(TV) For any agent *S*, and any proposition *p*, it is permissible that *S* believes *p* if and only if *p* is true.

The knowledge view entails,

(KV) For any agent *S*, and any proposition *p*, it is permissible that *S* believes *p* if and only if *S* knows that *p*.

(KV) is stronger than (TV). According to (KV), it is necessary but insufficient for the permissibility of a subject's believing a proposition that the proposition is true. So if it is permissible to believe that *p* according to (KV) then it is permissible to believe that *p* according to (TV). However, according to (TV), it is sufficient but unnecessary for the permissibility of a subject's believing a proposition that the subject knows the

proposition. So it does not follow from the permissibility of believing that *p* according to (TV) that it is permissible to believe that *p* according to (KV).

Some philosophers reject the claim that there is a fundamental norm, or standard of correctness, for belief. Many philosophers think there is, but not all do. We might call philosophers who endorse some version of the view that there is fundamental norm, or standard of correctness, for belief *doxastic absolutists*. We might call philosophers who deny that there is any fundamental norm, or standard of correctness, for belief *doxastic relativists*[5]. The case for doxastic relativism often appeals to examples in which there is no prima facie obligation to avoid false belief. Why, for example, is it wrong for someone with a terminal illness to go on believing that they are well when they have evidence to the contrary if this improves their quality of life; or for a parent to believe, contrary to the evidence, that their missing child is still alive if this belief is the only thing that keeps them going; or for a person to be in denial about a certain aspect of their past which it would be too difficult for them to openly confront? It might be objected that both (TV) and (KV) would appear to entail that in all such cases it is wrong for the individuals to have these beliefs.

I think proponents of (TV) or (KV) would reject the claim that their view entails that in the aforementioned examples it would be wrong for the individuals to have the beliefs in question. They would argue that claims about the fundamental norm, or standard of correctness, of belief are claims about the distinctively *epistemic* norm of belief. Furthermore, they would argue that norms are defeasible. They can be trumped by other considerations. Consider an example that Timothy Williamson gives in the context of discussing the norm of assertion. Williamson argues that one should believe a proposition only on the condition that one knows it. As I will explain in more detail below, Williamson also argues that one should assert a proposition only on the condition that one knows it. In defense of this latter claim, he writes,

"I shout, "That is your train," knowing that I do not know it is, because it probably is and you have only moments to catch it. Such cases do not show that the knowledge rule is not the rule of assertion; they merely show that it can be overridden by other norms not specific to assertion". (1996: 508)

[5] I am borrowing these general labels from Jose L. Zalabardo (2010). For a defense of doxastic relativism, see David Papineau (*forthcoming*).

Similarly, I think Williamson would argue that examples in which it seems completely unobjectionable for a person to believe something that she does not know do not show that knowledge is not the norm of belief. Rather, these are cases in which the norm of belief is defeated by other non-epistemic norms – i.e. by norms non-specific to belief.

4.3 Williamson's Argument for the Knowledge View

I will now present a positive argument for (KV) given by Timothy Williamson. Williamson's argument depends on his claim that knowledge is the fundamental norm, or "constitutive rule", of assertion. Williamson defines the constitutive rule of an act as a rule that is essential to the performance of that act. He does not mean that constitutive rules constitute the necessary conditions of the performance of acts. If one breaks the rule, it does not follow that one fails to perform the act. Rather, one's performance of the act is subject to criticism because one breaks the rule (2000: 240). Williamson defends,

> "(The knowledge rule) One must: assert p only if one knows p." (2000: 243)

I will call the knowledge rule for assertion (KR). Williamson contrasts (KR) with an alternative account of the norm of assertion,

> "(The truth rule) One must: assert p only if p is true." (2000: 242)

I will call the truth rule for assertion (TR). Williamson argues that there are strong reasons for favouring (KR) over (TR). He also suggests that if (KR) is true that we have reason to think that knowledge, rather than truth, is the norm of belief as well. He writes,

> "It is plausible…that occurently believing p stands to asserting p as the inner stands to the outer. If so, the knowledge rule for assertion corresponds to the norm that one should believe p only if one knows p". (2000: 255-6)

Admittedly, talk of occurently believing *p* standing to asserting *p* "as the inner stands to the outer" is not entirely perspicuous. Nonetheless, I take the general point to be that since assertion is the direct expression or linguistic manifestation of belief, if assertion is

governed by a knowledge norm then so must belief be. I think this is plausible. If so, we can infer from the norm of assertion to the norm of belief via the following linking principle:

1. For any agent *S*, it is permissible for *S* to assert that *p* if and only if *S* knows that *p*. *(knowledge norm of assertion)*
2. For any agent *S*, it is permissible for *S* to assert that *p* if and only if it is permissible for *S* to believe that *p*. *(linking principle)*
3. Therefore, for any agent *S*, it is permissible for *S* to believe that *p* if and only if *S* knows that *p*. *(KV)*

Williamson presents three main reasons for favouring (KR) over (TR). To begin with, he claims that (TR) and (KR) are intended to express the constitutive rules of assertion. Williamson says that such rules are essential to and individuating of the act in question. But he argues that (TR) fails to differentiate assertion from other kind of speech act. Williamson's examples are conjecturing and swearing (in the sense of making an oath). He says that it is good to conjecture the true and bad to conjecture the false. Yet assertion and conjecture are distinct types of act. The evidential norms governing conjecture are more relaxed than those governing assertion. If *p* is merely more probable on one's evidence than ~*p* then it would be acceptable to conjecture that *p*. However, it would not be acceptable to assert that *p*. Similarly, Williamson says that it is good to swear to what is true and bad to swear to what is false. But assertion and swearing are also different types of act. The evidential norms governing swearing are more stringent than those governing assertion. Swearing only seems acceptable if one has unusual grounds for certainty.

In fact, I do not find this initial argument very strong. It is not obvious why we should not state the constitutive rule of conjecture as follows:

> One must: conjecture that *p* only if *p* is at least as probable on one's evidence as ~*p*.

Admittedly, if one conjectured that *p* on the basis that it was at least as probable as ~*p*, but *p* turned out to be false, someone might reasonably object that one made the wrong

guess. However, it is not clear that in objecting that *p* was the wrong guess the objector is claiming that fault lies with the act, and not merely with the content. That is, it is not clear that the objector is not simply pointing out that *p* is false. Equally, it is not obvious to me why we should not state the constitutive rule of swearing as follows:

One must: swear that *p* only if one is certain that *p*.

This would enable us to hold onto (TR) whilst differentiating the act of assertion from that of conjecture and swearing.

A stronger argument that Williamson gives in favour of (KR) appeals to lottery-related considerations. Williamson says that assertion has an evidential norm. He says that it is better to make an assertion on the basis of adequate evidence than on the basis of inadequate evidence. If we assume (TR) then the evidential norm of assertion is derivative from (TR). In other words, one ought to have evidence for one's assertions because one ought to assert only the truth. Williamson states the principle underlying this derivation in terms of the following schema:

> "(I) If one must (ϕ only if p is true), then one should (ϕ only if one has evidence that p is true)". (2000: 245)

As it stands, (I) leaves room for a great deal of variation. The question is how much evidence for *p* is required for one to be justified in *ϕ*-ing. Williamson says that on a charitable reading of (I), the required weight of evidence will vary with the badness of *ϕ*-ing when *p* is false. He then asks whether (I) can explain the weight of evidence that speakers are required to have for their assertions in terms of the badness that we attribute to asserting something false. Williamson thinks it cannot. He gives the following example:

> "Suppose that you have bought a ticket in a very large lottery. Only one ticket wins. Although the draw has been held, the result has not yet been announced. In fact, your ticket did not win, but I have no inside information to that effect. On the merely probabilistic grounds that your ticket was only one of very many, I assert to you flat-out 'Your ticket did not win', without telling you my grounds. Intuitively, my grounds are quite inadequate for that outright unqualified assertion, even though one can construct the example to make its probability on my evidence as high as one likes, short of 1, by increasing the number of tickets in the lottery." (2000: 246)

Williamson argues that (I), and so (TR), cannot account for the fact that there are not adequate grounds for asserting, "Your ticket did not win", and so for the fact this assertion seems open to criticism, despite it's being exceedingly probable on the evidence that it is true. Williamson argues that (I) could only explain the weight of evidence required to justify making this assertion in terms of the extremely bad consequences of asserting it. However, in actual fact, in making this assertion, one runs a one-in-a-million risk of inflicting consequences of extremely limited badness. On the other hand, Williamson argues that (KR) does explain our intuitions about the assertion of such lottery propositions. According to (KR), since I clearly do not know that your ticket will lose it is wrong for me to assert that it will. Williamson points out that (KR) also accords with the fact that the natural way for one to articulate one's criticism of the assertion that a given ticket will not win is by saying, "You don't know that".

Williamson's final argument is that (KR) is supported by conversational patterns. Firstly, he notes that, excepting cases of assertion in which it is obvious to both speaker and listener how the assertor knows what she is asserting, it is normally appropriate to respond to an assertion with the question, "How do you know?" which presupposes that the assertor does in fact know, or is at least expected to know. (KR) explains why this presupposition is legitimate since it entails that failure to supply an answer to the question of how one knows implies an absence of warrant. Secondly, Williamson argues that (KR) is able to explain a version of Moore's paradox in which 'know' replaces 'believe'. According to the original paradox, there is something wrong with asserting, "*A* but I do not believe that *A*". But Williamson observes that it also sounds wrong to assert, "*A* but I do not know that *A*". (KR) easily explains what is wrong with both assertions. According to (KR), in order to have warrant to make either assertion one would have to know both conjuncts. However, one cannot know both *A* and that one does not believe that *A*. Nor can one know both *A* and that one does not know that *A*. In both cases, one's knowledge of the latter conjunct undermines one's knowledge of the former. It would seem that (TR) could explain the original paradox. Since (TR) states that one ought to assert something only if it is true, it would seem to imply that if one does not believe that something is true one should not assert it. Therefore, one should not assert, "*A* and I do not believe that *A*". However, Williamson argues that (TR) cannot explain the version of the paradox that replaces 'believe' with 'know'. This is

because (TR) implies that one should believe that *p* only if one has good evidence that *p* is true. However, having good evidence for something is compatible with not knowing it.

Jennifer Lackey rejects (KR) on the basis of purported counterexamples in which a person makes assertions that seem entirely acceptable, but in which the person does not know what she is asserting. For example, she considers the case of a science teacher who does not believe in evolutionary theory because she is a creationist, but who teaches evolutionary theory to her students nevertheless. Lackey argues that the teacher's assertions regarding evolutionary theory are completely appropriate, but she does not possess knowledge of her assertions because she does not believe them. In which case, Lackey argues, it would seem that (KR) is false (2007: 599). It seems to me that examples of this type do not pose a serious threat to (KR). As Jonathan Kvanvig (2010) argues, Williamson states that the norm of assertion is defeasible. But in the case described, there are other, non-epistemic factors that count in favour of the teacher's assertions and that explain why they are appropriate in a way that is consistent with (KR). There is overwhelming scientific evidence in favour of evolutionary theory and the teacher has certain responsibilities that accrue to the particular social role that she is paid to fulfil. I think that we might also question whether the teacher's purported assertions really are assertions at all. It might be argued instead that in an example such as this the teacher is not truly asserting, but is merely acting as a mouthpiece for the transmission of communally possessed knowledge.

I think that a more serious objection to (KR) Williamson himself anticipates. Towards the end of his discussion of assertion in *Knowledge and its Limits*, Williamson considers two alternative proposals to (KR). Firstly, he considers,

> "(The BK rule) One must: assert p only if one believes that one knows p." (2000: 260)

Williamson says that (BK) can explain the conversational phenomena that he cites as evidence for (KR). If one is committed to believing one knows one's assertions then the challenge "How do you know?" makes sense. Furthermore, one cannot both believe that one knows A and believe that one knows that one does not know A. Hence (BK) explains what is wrong with asserting "*A* and I do not know that *A*". Though

Williamson does not say this, it is also worth noting that it explains what is wrong with asserting "Your ticket did not win" in the lottery example, so long as we treat the example as assuming that the assertor believes that she does not know this. Nonetheless, Williamson rejects (BK) because he says it fails to account for the fact that there is something wrong with asserting irrational beliefs. Williamson imagines irrationally believing that one knows that G.E. Moore was a serial killer and asserting this. He argues that not only would there be something wrong with the assertor (i.e. the fact that the assertor has this irrational belief), but there would also be something wrong with the assertion itself. However, according to (BK), there is nothing wrong with the assertion.

Williamson considers a second alternative to (KR),

> "(The RBK rule) One must: assert p only if one rationally believes that one knows p." (2000: 261)

Williamson says that (RBK) improves on (BK) by eliminating its counterintuitive consequences. However, he raises four problems with it. Firstly, he considers a complicated, multi-premise argument that leads to a contradiction. He argues that since it might be rational for one to believe that one knows each premise of the argument, but not to believe one knows the conjunction since taken together they lead to paradox, (RBK) implies that warrant to assert is not closed under conjunction. Williamson says that this consequence would be "disturbing, but not clearly absurd" (2000: 261). Secondly, Williamson says that (RBK) cannot explain why intuitively there is something wrong with asserting propositions that it is rational for one to believe one knows, but which are false. Thirdly, Williamson says that since (KR) and (RBK) are supported by the same evidence, but (KR) is simpler, the onus of proof is on (RBK). Finally, Williamson objects that (RBK) makes it too easy for someone who lacks the authority to assert p to confer that authority on someone else, since one might know $\sim p$, but create sufficiently misleading appearances to make others rationally believe that they know p.

It strikes me that on the whole Williamson's objections to (BK) and (RBK) are not very strong. With regards to (BK), while there would be something wrong with irrationally believing that G.E. Moore was a serial killer, and there would (very probably) be something wrong with the content of the assertion expressing that belief, given that it is (very probably) false, it is not intuitively obvious that there would be anything wrong

with the act of making the assertion itself if the assertor truly believed that she knew it[6]. With regards to (RBK), it does not seem obviously 'disturbing' if warrant to assert was not closed under conjunction. It also does not seem obvious that there is anything wrong with asserting rational, false beliefs. Again, there may be something wrong with the content of such assertions, since they are false, but there is not obviously anything wrong with the act of making the assertion. The same could be said for Williamson's example of someone who has arrived at a rational, false belief through the deception of another – i.e. Williamson's fourth objection to (RBK). With respect to the third, it is not immediately transparent why Williamson thinks that (KR) is simpler that (RBK). (RBK) is not theoretically inelegant – it does not multiply hypotheses beyond necessity. Indeed, as I have suggested, some may have the intuition that there is nothing wrong with the act of asserting rational, false beliefs. (RBK) would explain this. (KR) cannot.

For my purposes, it does not matter which of (KR), (BK) and (RBK) is true. The objections that I will outline to cognitivist theories of intention in section 4.5 can be run using a formulation of the standard of correctness for belief corresponding to any of these principles. I return to this issue in 4.5.

4.4 Further Arguments for the Knowledge View

It might be thought that (KV) conflicts with the way we normally criticize and evaluate beliefs. In this section I will suggest that this is not so.

An obvious proposition that might be thought to present a counterexample to (KV) is something like, "I will not be killed in a road accident tomorrow". It might be argued that ordinarily one cannot know this, but, contrary to (KV), there is nothing wrong with believing it. Let us say that a person *outright* believes that *p* if she takes it to be settled that *p*. (KV) does imply that one should not outright believe that one will not be killed in a road accident tomorrow if one does not know this. However, it is open to the defender of (KV) to defend this position by appealing either to the notion of degrees of confidence, or by sticking to a coarse-grained conception of belief but maintaining that one's attitude ought to take the form of a probability judgement. So, given the statistical unlikelihood of being killed in a road accident tomorrow, the defender of (KV) might

[6] For this kind of response to Williamson on this point, see also Jessica Brown (2008: 93-4).

argue that, while one should not outright believe that one will not be killed in a road accident tomorrow, either one should have a high degree of confidence that one will not, or one should believe that it is highly probable that one will not[7].

Another supposed counterexample to (KV) is given by Dan Whiting (*forthcoming*):

"Suppose David asks, 'Who do you believe will win the next election?' Kelly might reply, 'The Republicans'. It would be very odd for David to reply, 'You don't know that!' And it would be entirely appropriate for Kelly to reject this challenge by saying, 'I never said that I did - I was only telling you what I believe'. Note that David might be right that Kelly does not know this but, still, his remark seems out of order."

My response to this example is that asking someone what she believes is, at least in the context under consideration, plausibly construed as asking her something more hedged or cautious then what she *outright* believes. This is why it would be strange to reply, "You don't know that!" What is implicitly being asked is what the person expects, or her judgement about what is likely. In asserting, "The Republicans", the interlocutor is stating her expectation, not what she outright believes. This is why the example reads like a trick. In responding, "You don't know that!" David is shifting the question from one that is implicitly about what Kelly expects to one about what Kelly outright or flatly believes. Suppose Kelly replied, "I *absolutely* believe that the Republicans will win", or "I *truly* believe the Republicans will win". In that case, David might well object, "But you don't know that". At the very least, David might query why she has full belief that the Republicans will win when she does not know. Intuitively, this would be a natural response even if both David and Kelly knew on the basis of evidence that it is probable that the Republicans will win, and knew that they both know that it is probable. Such responses presuppose that in having full or outright belief that *p* one is expected to know that *p*. (KV) makes sense of this.

A final sort of case that might be thought to create problems for (KV) involves testimony. We need to distinguish between two questions – whether testimony can be a basis for permissible or correct belief, and whether testimony can be a basis for justified belief. If testimony can be a source of knowledge, as is commonly assumed, then it follows from both (TV) and (KV) that it can be *permissible* to believe propositions on the

[7] This is Williamson's view (2000: 255).

basis of testimony. However, the fact remains that people communicate inaccurately, make mistakes and deceive. It might be thought that this raises a problem for the standard of justification implied by (KV). The standard of justification implied by (KV) is that belief is justified only if one has adequate grounds for taking oneself to know. It might be argued that the fact that people communicate inaccurately, make mistakes and deceive entails that we never have adequate grounds for taking ourselves to know things on the basis of testimony. In which case, it would seem that the standard of justification implied by (KV) is too demanding. However, it is not at all clear that the fact that people communicate inaccurately, make mistakes and deceive entails that we never have adequate grounds for taking ourselves to know things on the basis of testimony. Consider the parallel case of perception. Perceptual experience is subject to error and inaccuracy. This hardly shows that we are never justified in taking ourselves to know things on the basis of perceptual experience. There is only a problem here if one's take the position of a committed sceptic – namely, the fact that a procedure can sometimes issue in false belief is in and of itself a reason never to trust the procedure. Of course, we do sometimes lack grounds for taking ourselves to know things purely on the basis of testimony where there are special reasons for doubt, just as we sometimes lack grounds for taking ourselves to know things purely on the basis of perception where there are special reasons for doubt. A clear example would be receiving directions from a stranger in the street. In such a case, one may well lack grounds for taking oneself to know what one has been told – for example, perhaps the person one asked seemed hesitant, unsure or preoccupied. In such a case, (KV) implies that one should not *outright* believe the directions one has been given. Rather, one should have a degree of confidence in the accuracy of the directions, or form a probability judgement about their accuracy, correlative to the evidence available – i.e. the speech and behaviour of one's guide.

4.5 Objections to Weak and Strong Cognitivism

In this section I outline objections to both weak cognitivism and strong cognitivism. I begin with strong cognitivism. The following is an argument for the claim that if (KV) is true then the fundamental norm of belief cannot be the fundamental norm of intention. Therefore, given (EC), the claim that subjection to the norm is the core or essence of belief, intention cannot be a kind of belief. The argument employs the letters *M*, *S*, and *p* to refer schematically to propositional mental states, subjects and propositions

respectively. *M* is used both as a noun, to refer to a particular kind of propositional mental state, and as a verb, to refer to the taking of a particular kind of attitude towards a proposition:

1. For any propositional mental state *M*, necessarily (if *M* is any kind of belief then *M* is subject to the standard of correctness of belief). *(premise: (EC))*
2. For any subject *S*, and any proposition *p*, it is permissible that *S* believes that *p* if and only if S knows that *p*. *(premise: (KV) = the standard of correctness of belief)*
3. Therefore, for any subject *S*, any mental state *M*, and any proposition *p*, necessarily (if *M* is any kind of belief then it is permissible for *S* to *M* that *p* if and only if *S* knows that *p*). *(from (1) and (2))*

*

4. An intention to ϕ is a kind of belief with the propositional content, 'I will ϕ'. *(hypothesis to be tested: strong cognitivism)*
5. Therefore, from (3) and (4), for any subject *S*, it is permissible for *S* to intend to ϕ if and only if *S* knows that *S* will ϕ. *(Conclusion)*[8]

The argument is valid, but leads to a false conclusion. The reason that the conclusion is false is that the conditions on knowledge are just much more stringent than the conditions on permissible intending. Consider the following argument schema:

i. If *S* knows that *S* will ϕ then *S* is in a position to know that *E* will not occur.
ii. *S* is not in a position to know that *E* will not occur.
iii. Therefore, *S* does not know that *S* will ϕ[9].

It is very easy to come up with values for 'ϕ' and for '*E*' such that the fact that *S* fails to know that *E* will not occur defeats *S*'s knowledge that *S* will ϕ. To take a fairly

[8] The statement of the conclusion in (5) does not precisely match the language of (3) because statements of intention cannot naturally be given a propositional attitude construction. A less natural, but more transparent reading of (5) for the purposes of the argument, would be: 'from (3) and (4), for any agent *S*, it is permissible for *S* to intend that (I will ϕ) if and only if *S* knows that (I will ϕ). Michael Thompson has argued that the fact that expressions of intention cannot naturally be given a propositional attitude construction indicates that intentions are not propositional attitudes (2012: 127-8). This strikes me as certainly plausible. However, the cognitivist might argue that it is not obviously true. They might argue that the structure of surface grammar is simply misleading in this respect.

[9] This argument schema was suggested to me by a similar schema presented by Ian Phillips in his paper, 'Ignorance of the Future' (*in preparation*).

uncontroversial example, suppose that I intend to fly abroad in nine months time. Even if I have every reason to suspect that I will fly abroad in nine months time, it seems extremely plausible that I do not know that I will. I might fall ill, a pilot strike might be called, or any number of other contingencies might arise. People fail to catch their flight for one reason or another all the time. In modal language, there are many nearby possible worlds in which I do not get on that plane. So I could not 'safely' believe that I will get on that plane. Therefore, according to (5), I ought not to intend to fly abroad in nine months time. But this seems wrong. Whether or not I ought to have this intention depends on whether or not I rightly take myself to have good reason to go abroad in nine months time. Whether or not I know that I will go abroad seems irrelevant[10]. Such examples are multipliable, with the result that, were (5) true, many of our ordinary, everyday intentions would be objectionable.[11]

If a valid argument has a false conclusion at least one of the premises must be false. I will not argue explicitly for (1). I take it for granted that when discussing the fundamental norm, or standard of correctness, of belief philosophers are discussing a property such that if anything is a belief then it has the property of being subject to that characteristic norm. I believe that (2) is plausible and well motivated. (3) is an implication of prior premises. Therefore, I suggest that (4), strong cognitivism, should be rejected. Call this the *knowledge argument against strong cognitivism*.

A related argument can be constructed against weak cognitivist theories of intention. According to weak cognitivism, if an individual intends to perform some action then she believes that she will perform that action. In conjunction with (KV), it follows that

[10] A further, possible argument against (5) might appeal to the idea that there are no facts of the matter about statements expressing future-contingents. If statements expressing future-contingents lacked a truth-value then, since knowledge is factive, such statements could not be known. Assuming (5), it would follow that one could not permissibly intend to do anything. A number of philosophers have defended the view that we cannot possess knowledge of the future on different grounds. Such philosophers argue that possessing knowledge requires possessing some sort of 'source' – something such as perception, memory or deductive proof that guarantees truth. Again, assuming (5) it would follow that one cannot permissibly intend to do anything. For a recent defense of the view that we cannot possess knowledge of the future, see Phillips (*in preparation*).

[11] Note that it will not do to respond to such counterexamples by appealing to some sort of contextualist epistemology. For example, suppose that I intend to start cooking a spaghetti dinner at 6pm in exactly three weeks. It might be argued that in most contexts I do know this. So in most contexts it is permissible for me to have this intention. However, asking the question of whether I know shifts the epistemic standards upwards. The implication of this would be that the permissibility of intending would be equally context-relative. Simply by raising the challenge of whether I know one could render my intention impermissible. It seems to me that this is not an acceptable consequence. Unfortunately, I do not have the space here to engage in a fuller discussion of contextualism.

intending to do something that one does not know one will do entails impermissibly believing that one will do something that one does not know one will do. But it is implausible that intending to do something that one does not know one will do *entails* having an impermissible belief. It is implausible that having very ordinary, unobjectionable kinds of intentions should condemn one to a state of epistemic vice or defect. As I have already suggested, there are innumerable cases in which we form intentions to do things that we do not know that we will do where there seems nothing objectionable at all about our having the intention in question. The consequence of the conjunction of weak cognitivism and (KV) that, short of revising one's intention, in all such cases one is condemned to a state of epistemic vice or defect is not an acceptable one. Since the conjunction of (KV) with weak cognitivism has this implausible consequence one of the conjuncts should be rejected. Either (KV) must be false or weak cognitivism must be false. I believe that (KV) is plausible and well motivated. Therefore, I suggest that weak cognitivism should be rejected. Call this the *knowledge argument against weak cognitivism*.

At the end of section 4.3 I considered some alternatives to Williamson's (KR) rule for assertion that Williamson calls the (BK) rule and the (RBK) rule. Recall that (BK) states that one must assert a proposition only if one believes one knows it, whereas (RBK) states that one must assert a proposition only if one *rationally* believes one knows it. I suggested that Williamson fails to satisfactorily eliminate (BK) and (RBK). Further, I suggested that, for my purposes, it does not matter which of (KR), (BK) and (RBK) is true because the objections I outline to cognitivist theories of intention in this section can be run using a formulation of the standard of correctness for belief corresponding to any of these statements of the norm of assertion. Here is why. Returning to the example given above, suppose that not only do I not know that I will fly abroad in nine months time, but I (rationally) believe that I do not know this. If it is correct to assume that the norm of assertion corresponds to the norm of belief, any of these principles concerning the norm of assertion, in conjunction with a strong cognitivist analysis of intention and (EC), counter-intuitively implies that there is necessarily something objectionable about my intending to fly abroad in nine months time - or, for that matter, intending anything that I (rationally) believe that I do not know I will do. Equally, if it is correct to assume that the norm of assertion corresponds to the norm of belief, any of these principles, in conjunction with a weak cognitivist analysis of intention, counter-

intuitively implies that intending to do something that I (rationally) believe I do not know that I will do entails having the impermissible belief that I will do what I intend.[12]

4.6 Potential Replies to the Objections

The most obvious response to the arguments outlined in 4.5 would be to reject (KV). I cannot attempt a conclusive defense of (KV) within the space of a single chapter. Nonetheless, I have argued that (KV) is a plausible and well-motivated position, and that the objections that I have raised can be motivated even using the weaker variations of (KV) corresponding to the (BK) and (RBK) rules. I will now consider two potential replies that do not turn on the rejection of (KV). To begin with, recall that it was claimed that (KV) states the uniquely epistemic norm of belief and that norms are defeasible and can be overridden. In light of this, a strong cognitivist might argue that even if (KV) is true it does not follow from the conjunction of (KV) and the strong cognitivist identification of intention with a kind of belief that *all things considered* it is permissible to intend to ϕ if and only if one knows that one will ϕ. For it may be the case that it is in the nature of intention that in every case of intending to perform some action where one lacks knowledge of what is intended there are countervailing considerations that override the knowledge norm. Call this the *defeasibility reply from strong cognitivism*. Similarly, a weak cognitivist might argue that even if (KV) is true it does not follow from the conjunction of (KV) and the weak cognitivist claim that intending to ϕ necessarily entails believing that one will ϕ that intending to do something that one does not know one will do entails having an impermissible belief. For the weak cognitivist

[12] Arguably a version of the objections could be run even using (TV). The conjunction of cognitivism with (TV) counter-intuitively implies that there is necessarily something wrong or objectionable about intending to perform an action when it is false that one will perform that action. I think there is an argument to be made for thinking that even this weaker criterion of correctness is too stringent compared to the standards of practically rational intending. It seems to me that it is not necessarily the case that one ought not to have intended to do something where it turns out that one did not succeed. However, in my view, such an argument would be less compelling than the objections that I have formulated using (KV). This is because often we intend to do things that we not only do not know we will do, but that we know we do not know we will do. The conjunction of cognitivism with (KV) entails that this is not only incorrect, but one would be open to criticism for this. For in such cases there would be no failure of transparency that excuses the violation of the norm. On the other hand, if one has a 'false intention' this will only be something that could be discovered retrospectively, after the fact. Therefore, assuming that one has suitable grounds for forming the intention-belief - which one would if, say, intentions were causally self-fulfilling beliefs that represent themselves as such (see chapter three) - the cognitivist could argue that while it is true that having false intentions is in some sense incorrect or objectionable, one would not be open to criticism for having them. Thus the mere impermissibility of having false intention-beliefs, while perhaps counter-intuitive, would not make any substantive difference to our everyday practices of forming and assessing intentions.

might also propose that it is in the nature of intention that in every case of intending to perform some action where one lacks knowledge of what is intended there are countervailing considerations that override the knowledge norm. Call this the *defeasibility reply from weak cognitivism*. My response to this first line of counterargument is that there has to be some positive reason for claiming that in every case of intending perform some action where knowledge of what is intended is lacking the knowledge norm if defeated. However, it seems to me that there is no positive rationale for making this claim. Suppose the reason given is that in every case of permissibly intending to perform some action where knowledge of what is intended is lacking the knowledge norm is defeated by the benefits of committing to doing what one has good reason to do. In this case I suggest there is a dilemma. If the defender of the defeasibility reply from strong cognitivism presses this response it becomes unclear why we should say that knowledge is the fundamental norm or standard of correctness of intention in the first place. If in every case of forming an intention the norm is overridden then why say that intention is subject to the norm at all. Equally, if the defender of the defeasibility reply from weak cognitivism presses this response it becomes unclear why the should say the attitude purportedly involved in or entailed by intention is subject to the knowledge norm. Therefore, it becomes unclear why we should think that intention involves or entails belief.

In order to get to grips with the second potential line of response to the knowledge arguments against strong and weak cognitivism it will help to return to a standard objection that has been made against cognitivist theories of intention, which I discussed in chapter one when examining Bratman's objections to the view that an intention to ϕ is reducible to the combination of the predominant desire to ϕ and the belief that one will ϕ because one so desires. Bratman's third objection to this type of reductionist analysis was that in many cases an agent might intend to do something, but be agnostic about whether she will actually do that thing. Such cases are purported to constitute counterexamples to the cognitivist's claim that intending to act is constitutively related to believing that one will so act. Bratman considers two kinds of scenarios in which this appears to be the case (1987: 37-41). In the first instance, he says that a person might be agnostic about whether she will do what she intends because she is unsure she will succeed in her endeavour. For example, he says that a person might intend to move a log blocking her driveway but be unsure whether she will manage to move the log.

Secondly, he argues that a person might intend to do something whilst being agnostic about whether she will even try to do that thing. Bratman's example is of someone who intends to stop off at the bookstore but is unsure whether she will because she believes that there is a significant possibility that she will forget. Recall that one way for the cognitivist to respond to the first sort of case is to re-describe it in a way that is compatible with thinking that intending to do something necessarily involves believing one will do it by importing the language of 'trying'. So it might be argued that the individual does not intend to move the log, but intends to *try* to move the log. Plausibly, this would be something that she believes. This kind of response is perhaps not so obviously available in Bratman's second example, which is deliberately designed to eschew that sort of reply. Nonetheless, in this second example, it might be argued that the agent's intention is conditional in some way. For instance, it might be argued that she intends to stop off at the bookstore *if all goes to plan*, or she intends to stop off at the bookstore *on the condition that she remembers*. Once again, plausibly, this would be something that she believes.

In chapter one it was noted that one thing that seems paradoxical about this strategy of re-describing the agent's intention in terms of trying, or as conditional or qualified in some way, is that while we might well expect the agent to say, for instance, "I intend to try to move the log", in describing what her intention is, it also seems correct to say of the agent that her intention is to move the log. So it seems true to say both that the agent intends to ϕ, *simpliciter*, and that she intends to try to ϕ. How can we explain this? As already mentioned, Velleman argues that sometimes we use the terms 'intention' or 'intend' to refer to an agent's 'goal'. He says that an agent's goal is what she is ultimately motivated to do. It is some achievement which she acts 'with the intention' of bringing about. On the other hand, Velleman says that sometimes we use the terms 'intention' and 'intend' to refer to whatever the agent has decided upon doing at the close of deliberation. Velleman glosses this by saying that sometimes we use the terms to refer to what he calls the agent's 'goal-states', whereas at other times we use them to refer to what he calls the agent's 'plan-states' (1989: 112). According to Velleman, the concept of 'intention' is fundamentally polysemous. The terms 'intention' and 'intend' are simply used to refer to different things. In the kinds of counterexamples to cognitivism typically given, Velleman says that in saying that the agent intends to ϕ (e.g. "move the log"), we a referring to the agent's goal-state. And in saying that the agent intends to try to ϕ ("try to

move the log"), we are referring to the agent's plan-state. Velleman says that he is concerned with offering an analysis of the latter sense of intention. That is, he concerned with the sense of 'intention' that refers to what course of action an agent has decided upon doing.

Cognitivists might respond to purported counterexamples in which it appears that an agent intends to ϕ, but is agnostic about whether she will ϕ, by re-describing the agent's intention as being conditional in some way, or as being an intention to try to ϕ. In a similar vein, cognitivists might respond to the objections outlined above, not by attempting to reject (KV), but by arguing that if one does not know that one will succeed in doing something then one should not have an outright intention to do it. Rather, one should only form a conditional or qualified intention that would plausibly count as something that one knows. For instance, in the example given in section 4.5, it might be argued that the content of my intention ought to be of the form, 'I will fly abroad in nine months, if all goes to plan', or 'I will fly abroad in nine months time, if nothing prevents me and I don't change my mind. The strong cognitivist will argue that such an intention-belief does not violate (KV), whereas the weak cognitivist will argue that such an intention does not entail having a belief that violates (KV). In this way, cognitivists might attempt to reconcile their claims about intention with the view that knowledge is the norm of belief. Call this the *conditionalizing reply* to the knowledge arguments against strong and weak cognitivism.

My initial response to the conditionalizing reply is that, as already suggested, it seems plausible that the aim or point of intending to do something, unlike believing that one will do it, is not to secure knowledge of how one will act in the future, but to determine how one will act in the future. Assuming that I have good reason to fly abroad in nine months time, and assuming that there is not any special reason for me to expect that I will not (e.g. strikes are not expected for the time I am flying), there is no reason for me to form a merely conditional or qualified intention to go abroad in nine months time. But if (KV) is true, then there is a reason for me to only have a qualified or conditional belief that I will go abroad since I do not know that I will. However, at this point, proponents of cognitivism might charge me with begging the question against their theories. This is particularly true of a theorist like Velleman who sees an intimate

connection between intention and knowing what one is going to do. Here then is a further objection to the conditionalizing reply.

Suppose that I have some conditional intention – for example, "I'll go to the bookstore *if I remember*". Thus my intention is of the form, 'If condition *C* then I will ϕ'. Now suppose that I do *not* remember. The antecedent of my conditional intention is false. According to the standard, truth-functionalist account of indicative conditionals in terms of material implication, if the antecedent of a conditional is false the conditional as a whole remains true. So, even though I forgot to go to the bookstore, the content of my intention remains true. In which case, my intention is successful regardless. But this seems incredibly counter-intuitive. If someone subsequently asked me if I did what I intended, I am hardly likely to answer in the affirmative. The problem is re-enforced when we recall that strong cognitivism, in conjunction with (KV) and (EC), entails that whenever one permissibly intends to do something one knows that one will do it. It follows from this that whenever one permissibly intends to do something one cannot but succeed in carrying off one's intention. This seems like an absurd consequence of the view.

There are two replies that the conditionalist could make to this objection to the conditionalizing reply. Firstly, the conditionalist might appeal to a different account of conditionals. Non-truth-functionalists about conditionals argue that it not plausible to suppose that the falsity of the antecedent of a conditional entails the truth of the conditional as a whole. They argue that it cannot be right that the falsity of the antecedent entails the truth of, 'if *p* then *q*', whatever the values are for '*p*' and '*q*'. Rather, if the antecedent of a conditional is false then the conditional as a whole might be either true or false. For example, on Robert Stalnaker's (1968) account of conditionals, whether or not the conditional is true depends on whether in a nearby possible world, a world that is minimally different from the actual world, were the antecedent true then the conditional would be true. Now it might be argued that on a non-truth-functionalist understanding of conditionals, if the antecedent of a conditional intention of the form, 'If condition *C* is met, then I will ϕ', turns out to be false, it does not necessarily follow that the content of the intention remains true. Therefore, it does not necessarily follow that the intention is successful. However, it seems to me that ultimately this is not going to be of much help to the defender of the conditionalizing

reply. Supposing we adopt a non-truth-functionalist analysis of conditionals, it follows that if the antecedent of a conditional intention is false then there are two possibilities. One is that the content of the intention remains true. However, as I have already argued, this cannot be correct. The other is that the content of the intention is false. In which case, it is not an object of knowledge. Therefore, it follows from the conjunction of (KV) and cognitivism that the intention is impermissible. Consequently, irrespective of what account of conditionals we adopt, we still get the result that whenever one permissibly intends to do something one cannot but succeed.

At this point, the proponent of the conditionalizing reply might respond by re-introducing Velleman's distinction between plan-states and goal-states. She might argue that whenever one permissibly intends to do something one is guaranteed to succeed in what one intends qua plan-state, or in what one decided or settled on doing, but it does not follow that one is guaranteed not to fail in what one intends qua goal-state, or in what one is aiming to do. In this way, the conditionalist might invoke a distinction between different senses of success. But even the claim that whenever one permissibly has an intention one is guaranteed not to fail in what one decided on doing seems highly counter-intuitive. It is much more plausible to suppose that sometimes we just fail to go through on the choices that we make. It seems to me that, unless one was already motivated by prior theoretical commitments, few would want to accept this analysis.

I will end by considering one last possible response that might be given to the objections above. Suppose we accept the objections. Does it follow that we must abandon cognitivism? As Bratman defines them, 'cognitivists' are philosophers who claim that intending necessarily involves believing that one will do what one intends. However, it might be argued that on a broader conception of 'cognitivist' we might suppose that a philosopher is cognitivist just in case she posits a constitutive relation between having an intention and having *some sort* of cognitive attitude. In other words, a theorist of intention is a cognitivist just in case she argues that the state of intention necessarily involves some sort of attitude with a truth-evaluable content. If this is correct then why could a cognitivist not argue that intention necessarily involves, not belief, but some other cognitive state or attitude that is not subject to such stringent conditions of correctness? A cognitivism of this sort would be untouched by the objections raised in 4.5 above. For example, could a cognitivist, or perhaps 'quasi-cognitivist', not argue that

intending to perform an action necessarily involves, not believing that one will perform that action, but *accepting* that one will perform that action. Accepting a proposition involves acting or reasoning as one would if one knew that proposition. The functional role of acceptance is to guide thought and behaviour in lieu or in the absence of knowledge or belief. We might illustrate this with the example from Ginet presented in the previous chapter. Suppose that I have embarked on a trip, but I am uncertain whether I remembered to lock my front door. I expect that I did since locking the front door after leaving the house is my habit. Nonetheless I know that on this particular occasion I was slightly preoccupied and I have no clear memory of locking the front door. One possibility would be for me to turn back and check, but this would be very inconvenient. On the other hand, I do not have the time or the resources to continue thinking indefinitely about the question of whether I locked the front door. In this situation, I might simply decide to accept that I remembered to lock the front door. In accepting that I remembered to lock the front door I treat it as a fact that I locked it and so as a reason to continue on my way rather than turning back to check. I act as I would if I actually knew that I locked the front door. Accepting a proposition has a characteristic voluntariness that is not present in the case of belief. Accepting a proposition is something that one can simply decide to do. Another important difference between acceptance and belief is that acceptance, unlike belief, is not subject to a norm of truth or knowledge. A proposition need not be true or known in order for it to be correct for one to accept it. So one need not possess sufficient evidence for believing the proposition, or for taking oneself to know it, in order to be justified in accepting it. Acceptance is an attitude that is governed by non-epistemic norms. In both of these respects, the attitude of acceptance seems better placed than belief to figure in an analysis of intention. Furthermore, acceptance, like belief, seems to be subject to rules of conjunction and entailment. In which case, perhaps the thesis that intending to act necessarily involves accepting that one will so act could explain the norms of intention-consistency and means-end coherence governing intention by way of the involvement of acceptance in intention, instead of belief.

Despite these advantages over the thesis that intending to ϕ necessarily involves believing that one will ϕ, it seems to me that the idea that intending to ϕ constitutively involves accepting that one will ϕ cannot be correct either. This is because, as Bratman (1992) has argued, intention is a 'context-dependent' attitude, by which he means that whether it is

reasonable to have the attitude depends on context-relative practical considerations. Bratman defends the thesis that acceptance is a context dependent attitude by presenting a number of examples. One example that he considers involves an individual who is planning a construction project (1992: 6). Within the context of planning the project, she accepts that the total costs of the project will be at the top end of the estimated range, even though at present she can only get an estimate of the potential costs and so is unsure about what the actual total will be. The rationale for accepting that the total costs will be at the top end is that there is an asymmetry in the costs of her being mistaken in her estimate. If she underestimates the costs of the project then she runs the risk of running out of money. This would jeopardize the completion of the project. If she overestimates the cost of the project then while this might cause some unnecessary inconvenience, she does not run the risk of running out of money. In a different context, such as a situation in which she was betting with someone about what the total cost of the project will be, there would not be the same asymmetry in the costs of error. Whether she overestimated or underestimated the total costs, what she would stand to lose in the bet would be the same. Therefore, it would not be reasonable to simply accept that the total costs of the project will be at the top end of the estimated range. The problem is, as Bratman himself notes, intention is not context-dependent in the way in which acceptance is (1992: 13). In this respect, intention seems more like belief. For example, returning to the original example from the previous section, suppose that I intend to fly abroad in nine months time. I have no particular reason to doubt that I will fly abroad in nine months time. Therefore, in most contexts I take it for granted that this is what I will do. Nonetheless, I do not take myself to know that I will fly abroad in nine months time. After all, I am aware that there are various contingencies that might arise that could prevent this from happening. Consequently, there may be contexts in which I do not take it for granted – for example, we might imagine a situation in which I am offered a low price for insuring my plane ticket, which I accept on the basis that my flight might be cancelled, or I might miss my plane, or I might fall ill. However, even though in such contexts I do not accept that I will fly abroad in nine months time, I still intend to fly abroad in nine months time. Therefore, it cannot be the case that accepting that one will do something is necessary for intending to do it.

4.7 Conclusion

The aim of this chapter has been to trace the consequences of a widely respected view in epistemology for the theory of intention. I have argued that the view that knowledge is the norm of belief is incompatible with cognitivist theories of intention, both weak and strong. Strong cognitivism, in conjunction with (KV), implausibly entails that there is something objectionable about intending to do something that one does not know one will do. Weak cognitivism, in conjunction with (KV), implausibly entails that intending to do something that one does not know that one will do entails having an impermissible belief, and so condemns one to a state of epistemic vice or defect. I have argued that (KV) is plausible and well motivated. There are good reasons for thinking that it is true and, hence, for rejecting cognitivism. I hope to have shown that (KV) has wider implications than has hitherto been recognized.

Chapter 5

Intentional Action and Causal Deviance

5. Introduction

In previous chapters I examined two different approaches to understanding what sort of mental state an intention to act is. According to the planning theory, intentions are distinct and irreducible states or attitudes, non-analyzable in terms of any other supposedly more basic states or attitudes, and which are metaphysically bound up with planning. According to cognitivists, intending to perform an action is constitutively tied to believing that one will perform that action. Strong cognitivists argue that intention just is a kind of belief about one's future actions. Weak cognitivists argue that intending merely entails believing that one will do what one intends. My conclusion thus far is that intention is an irreducible mental state. It is not analyzable in terms of any other, supposedly more basic mental states. Even the weak cognitivist's thesis that intending necessarily entails believing that one will do as one intends is false. In this respect, I agree with the planning theorist. However, I reject the planning theorist's claim that intention is metaphysically bound up with planning. I reject the claim that to be the sort of creature that forms intentions just is to be a planning agent. Though my aim is to defend the irreducibility of intention, it is helpful to distinguish between two different sorts of irreducibility claim that might be made. On the one hand, there is what might be called the claim of 'weak primitivism'. According to this claim, intention is not reducible to any other, supposedly more basic, folk-psychological states or attitudes, or combination thereof, such as desire and belief. On the other hand, there is what might be called the claim of 'strong primitivism'. According to this claim, intentions are not only weakly primitive, but they are also conceptually primitive in the sense that there is nothing illuminating to be said about what they are. My aim is to defend the weak primitivism of intention. I do believe that there are interesting things to be said about what intentions are. In this respect, there is a certain parallel with claims that Williamson (2000) makes about the nature of knowledge. According to Williamson, knowledge is a basic mental state, rather than something to be analyzed in terms of belief, truth and certain other conditions. However, for Williamson, this does not mean that there is nothing interesting to say about knowledge, or that we cannot use the concept to shed light on the nature of other concepts, such as evidence, evidential probability and

assertion. In this chapter I argue that intentions are a particular kind of disposition – the disposition of an agent to pursue an aim or goal – and that this explains how intentions are causally related to, or control, actions. This proposal also fits comfortably with a claim made in previous chapters that we should think of intention as a state that has an aim – the aim of self-control – and it is this that explains the normative features of intention.

In order to motivate the thesis that intentions are a kind of disposition, in this chapter I consider the relation between intention and intentional action. In particular, I defend the following:

> *The Causal Theory*: For any agent *S*, and any instance of a type of behaviour, ϕ-ing, *S* ϕ's intentionally if and only if *S*'s ϕ-ing is caused in the right sort of way by *S*'s having an intention with an appropriate content.

The causal theory, as I understand it, is a view specifically about the nature of *intentional* action. It is compatible with thinking that the class of actions is not exhausted by the class of the intentional and hence that agents may perform actions that are not the causal consequence of a prior state of intending, or any other particular kind of antecedent mental state or attitude. Thus I see no inconsistency between the causal theory and acknowledging the widespread existence of what Brian O'Shaughnessy (1980) refers to as 'sub-intentional actions', such as absent-minded head-scratching, finger-drumming or leg-folding. It seems highly plausible that such sub-intentional bodily movements are instances of action even though they could not naturally or appropriately be described as intentional. Similarly, I see no inconsistency between the causal theory and acknowledging the widespread existence of various kinds of expressive actions – gestures and expressions such as smiling, jumping for joy or baring one's teeth in anger that are associated with emotion and that are neither reflexes nor exactly intentional[1]. Such behaviours are plausibly viewed as actions, but actions that manifest states or attitudes other than intention. In my view, the question of what makes an episode of bodily movement an action is clearly distinct from and more basic than the question of what

[1] For an interesting discussion of such expressive behaviour, see Goldie (2000).

makes behaviour intentional. In do not believe that the causal theory provides a plausible answer to that more basic question[2].

In stating the causal theory, I use the phrase, 'an intention with an appropriate content', instead of simply, 'an intention to ϕ', because not all philosophers who think that a given episode of behaviour is intentional in virtue of involving or being in some way appropriately related to the mental state of intention also accept what Bratman labels the 'Simple View' (1987: 111) - the view that for a person to intentionally ϕ she must intend to ϕ. Bratman rejects the Simple View on the grounds that there are cases in which it is perfectly rational for a person to intend to do both *A* and *B* but knows that she cannot do both. For example, he says that it would be perfectly rational for a person playing a computer game to guide a missile towards one target and simultaneously guide a missile towards a second target, even though she knows that if she is about to hit both targets the game will automatically shut down, because this will maximize her chances of hitting one of the targets and winning a reward. If she does *A* (e.g. hits target one) she will do so intentionally and if she does *B* (e.g. hits target two) she will do so intentionally. Therefore, according to the Simple View, she intends to do both *A* and *B*. However, Bratman claims that this would entail that the agent's intentions are strongly inconsistent with her beliefs and, therefore, that she is not being rational (1987: 114/5). Bratman argues instead that in such cases the agent's state is better characterized as one of intending to try to do *A* and to try to do *B*. He says that when a person intends to try to ϕ she acts 'with the intention' of ϕ-ing, but that the phrase, 'acts with the intention' is ambiguous (1987: 128/9). It can mean either that the agent is acting with the *further* intention of ϕ-ing, or that the agent is 'endeavouring' to ϕ, where this entails that the agent has a 'guiding desire' to ϕ (1987: 137). According to Bratman, if a person endeavours to ϕ then, if her endeavouring is successful, she ϕ's intentionally. However, endeavouring to ϕ does not entail intending to ϕ.[3]

The causal theory articulates the commonsense thought that when an agent acts intentionally her state of intending to act plays a causal role in the generation of her

[2] I believe that at least a necessary condition of some episode of bodily movement constituting an action is that the agent had the power to refrain from it. For a compelling defense of this thesis, see Steward (2012).

[3] For a criticism of Bratman's rejection of the Simple View, see McCann (1991).

behaviour. It also offers a non-mysterious way of understanding the relation between reasons and actions. Thus if we say, 'She drew the curtains because she intended to look outside', the causal theory proposes that we understand the 'because' here as expressing a causal relation between the agent's state of intention and her action. Nonetheless, like the causal theories of knowledge and perception, a long-standing objection to the causal theory of intentional action relates to the problem of so-called deviant causal chains. These are cases in which the occurrence of some type of behaviour or event seems to be caused by an agent's having an intention with an appropriately matching content, but where intuitively the occurrence of the behaviour or event does not qualify as intentional because the manner in which it was brought about was in some sense 'deviant' or 'wayward'. The challenge posed for the causal theorist is to come up with an adequate account of the basis of the distinction between deviance and nondeviance, and so to state precisely what the difference is between wayward causation and causation in 'the right way'. A number of philosophers have suggested that no adequate account will be forthcoming and have taken the problem of deviant causal chains to be a conclusive objection to causal theory of intentional action. For example, George Wilson writes that the problem of deviant causal chains "points to more than an infelicity or incompleteness in the various causalist proposals – it points, that is, to a global breakdown in the whole project of reduction" (1989: 258)[4].

My aim in this chapter is to defend the view that explanation of an agent's intentional actions appealing to an agent's intentions is a variety of causal explanation. However, while I think that explanations of intentional action appealing to intention represent one important and distinctive kind of causal explanation, I believe that they are not causal in the sense that has standardly been presupposed by causal theorists. My defense of the causal theory appeals to an argument recently made by John Hyman in a chapter of his forthcoming book, *Action, Knowledge and Will*. Hyman's aim in that chapter is to defend the view that intentional action is caused by desire. Hyman argues that the possibility of deviant causal chains only remains problematic for the view that intentional action is caused by desire on a 'Humean' conception of causality. He argues that we should understand the causal relation between desire and intentional action, not in 'Humean',

[4] See also Anthony Kenny (1975: 121). For some more recent pessimists, see Lily O'brien (2012); Lucy O'brien (2007: ch. 8); Scott R. Sehon (1997) and Helen Steward (2012). Donald Davidson was himself pessimistic about reaching a fully adequate account of sufficient event-casual conditions of intentional action. In his essay, 'Intending', he describes his account of intentional action as "incomplete and unsatisfactory" (1980: 87).

event-causal terms, but as the manifestation or exercise of a disposition. Hyman operates with a very particular conception of desire and one that one might think sounds very close to intention. I propose that we treat Hyman's analysis of the causal relation between desire and intentional action as a model for thinking about the relation between intention and intentional action. In section 5.1 I attempt to motivate a degree of pessimism about the prospects of reaching an adequate event-causal analysis of the distinction between deviance and nondeviance. Since it would not be possible to review all the literature in a single chapter, I focus on a particularly comprehensive treatment of the issues by John Bishop in his book, *Natural Agency* (1989). In section 5.2 I present Hyman's argument and attempt to articulate what I believe is correct in it. In section 5.3 I defend a crucial claim that Hyman makes, but does not defend or expand on, that dispositions are causally relevant to what they manifest. Finally, in section 5.4 I address two further worries that some may have about the causal theory of intentional action, aside from the problem of causal deviance.

5.1 Bishop and the Problem of Causal Deviance

In this section I consider, and ultimately reject, an attempt by John Bishop to offer satisfactory criteria for distinguishing between deviant and nondeviant causation. It will be helpful to begin by distinguishing between two different kinds of deviance: 'antecedential' deviance, on the one hand, and 'consequential' deviance, on the other[5]. In examples of antecedential deviance the deviance is internal to the agent because it affects the path between the relevant mental state or attitude and the initial bodily movement. The following is a classic example of antecedential deviance from Donald Davidson,

> "A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, " (1980: 79).

The example is set up as a counterexample to Davidson's theory of intentional action. According to Davidson's account (1980: Essay 1), an action *A* is intentional under a description *d* if and only if it was caused by a 'primary reason', which consists in the

[5] The terms 'antecedential' and 'consequential' were first used to demarcate between two kinds of deviance by Myles Brand (1984: 18).

pairing of a desire (or more broadly speaking, a 'pro-attitude') and a belief that rationalize the performance of *A* under the description *d*. Note, however, that even if we believe that the state or attitude of intending is distinct from and non-reducible to a Davidsonian primary reason, it could just as well be the climber's awareness of her irreducible and sui generis attitude of intending to let go of the rope that caused her to become nervous, in turn causing her to tremble and let go of the rope.

In cases of consequential deviance the deviance is external to the agent in that it occurs between the initial bodily movement and the ensuing consequences. The following is an example attributed to Daniel Bennett, but also quoted from Davidson,

"A man may try to kill someone by shooting at him. Suppose the killer misses his victim by a mile, but the shot stampedes a herd of wild pigs that trample the intended victim to death." (1980: 78)

Bishop offers separate and distinct accounts of the basis of the distinction between deviance and nondeviance in cases of antecedential and consequential deviance. I begin with his treatment of consequential deviance. Examples of consequential deviance involve deviance in what Bishop refers to as intended "nonbasic action". Bishop says that intending a nonbasic action involves intending some goal or end, *e*, where *e* is something that can only be brought about by the agent indirectly, by bringing about other events or states of affairs as a means to it. Bishop stipulates that in order to perform a nonbasic action, an agent must form and implement an "action plan" (1989: 128). In other words, the agent must reason towards and carry out suitable means to achieving *e*. Bishop then argues that an event or outcome constitutes an intentional nonbasic action if and only if the way it was brought about 'matches' the agent's action plan. In cases of consequential deviance, he claims that the reason why the event or outcome, though intended by the agent, does not count as intentional is that this matching condition is not met. Bishop stipulates a number of conditions that must be fulfilled in order for an outcome or event to match an agent's action plan. He begins by presenting the following three:

"First, the agent must perform the basic acts specified in the plan. Second, the agent's doing so must be intentional under the description given in the plan. And, third, the agent must perform these acts with the intention of attaining the end the plan was formed to serve" (1989: 130).

Bishop maintains that these conditions can account for why the outcome is not intentional in a number of examples of consequential deviance. However, he says that by themselves they are not sufficient to account for all cases. One such case is the stampede example quoted from Davidson. Bishops notes that in this example, the agent performs the basic acts specified in his plan – namely, aiming and firing the gun. These acts are intentional under the description given in his plan. They are also performed with the intention of attaining the end the plan was formed to serve. Nonetheless, the victim's death is not intentional. For this reason, Bishop argues that a fourth condition is required:

"the actual outcome must conform to the agent's beliefs (formed in his or her practical reasoning) about how it would be that the basic actions planned would yield the desired goal." (1989: 131)

However, the problem is that while the previous three conditions are too weak to exclude all cases of nonbasic deviance, the addition of this fourth condition renders Bishop's matching condition too strong. As Bishop himself observes, it rules out as unintentional examples of nonbasic action that do seem to be genuinely intentional. Bishop considers the example of a fastidious assassin who plans to kill a tyrant with a shot through the heart, but who fails to hit their heart and kills the tyrant with a headshot instead. The assassin kills the tyrant intentionally. However, the outcome was not brought about in a way that conforms to the assassin's beliefs about how her basic actions would yield the desired goal. It seems then that Bishop needs something weaker than the requirement that the actual outcome must conform to the agent's beliefs about how it would be that her basic actions would yield the desired goal. He says the following:

"The correct hint here, I think, is that what matters is whether the actual causal sequences is of a type that the agent could have considered in reasoning toward his or her action-plan…Given plausible background assumptions, the pig-stampeder could not have considered what actually occurred as a possible means of dispatching his enemy, whereas our assassin could certainly have considered aiming at the tyrant's head and could also have recognized that shooting at the tyrant's heart might have resulted in his getting a bullet through the head" (1989: 132).

However, this does not seem adequate either. Firstly, it is not obvious why the individual in the stampede example could not have considered the chain of events that actually occurred as a possible means of dispatching her enemy when reasoning about

how to achieve her aim. It seems fairly easy to fill in the details of the example in such a way that the agent could plausibly have considered this in her reasoning. Intuitively this would still be a clear case of consequential deviance. Secondly, it seems possible to imagine genuine cases of intentional nonbasic action in which the causal sequence running from the intended basic action to the intended outcome is not of a kind that the agent could have considered in his or her reasoning. These will be cases in which an agent instigates through her basic actions a process that she knows to reliably lead to some intended outcome, but where the nature of the process is inaccessible to the agent – for example, a young child operating a television remote control. So the requirement that the actual casual sequence be of a type that the agent could have considered in her reasoning does not even seem necessary for intentional nonbasic action.

This last example suggests taking a different tack. An alternative diagnosis of the source of the deviance in the stampede example might appeal to the notion of reliability. It might be argued that the reason why the death of the victim in the stampede example does not count as intentional is that the causal pathway leading from the agent's basic actions to the occurrence of the intended outcome was not a reliable one. David Pears (1975) defends an account of the distinction between consequential deviance and nondeviance that incorporates both a matching condition and a reliability condition. Pears makes the following three stipulations about an agent who does A intentionally:

> "(1) He must bring about A with the specific intention of bringing A about.
> (2) If A is non-basic, he must bring it about by reliable stages.
> (3)These stages must match their specifications in his plan." (1975:51)

Pears refers to the second and third conditions as 'primary' and 'secondary' reliability respectively (1975: 52). According to Pears, "A performance is primarily reliable if the goal is achieved by dependable stages". On the other hand, "a performance is secondarily reliable if its prior stages match their specification in the agent's plan" (1975: 52). Pears acknowledges that there are cases of intentional nonbasic actions which are secondarily reliable, but lack primary reliability. An example he gives involves a gunman who kills his enemy with the specific intention of hitting him with a ricochet (1975: 52). Equally, Pears acknowledges that there are cases of intentional nonbasic actions which are primarily reliable, but lack secondary reliability. For example, he cites the case of a gunman who unintentionally aims off to the left, but a crosswind blows her bullet on a

curved trajectory straight into her intended victims heart (1975: 53). In this case the intended outcome is achieved by dependable stages. However, there are two stages at which there is a mismatch with the agent's plan. Pears admits that in such cases it may indeed be correct to say that the agent brought about the outcome intentionally. Nonetheless, he maintains that there is a need to qualify this and that we cannot say that the agent achieved her goal intentionally *period* (1975: 53). Rather, we will have to add a qualifier such as, "but she was lucky". By contrast, in cases in which the intended outcome was brought about in a manner that lacked both primary and secondary reliability, such as the stampede example quoted from Davidson, Pears holds that "we are in real doubt whether to qualify the statement that he killed him intentionally or to deny it outright". In fact, contrary to Pears, it seems to me that there are clear cases of intentional nonbasic action in which the agent intentionally brings about an intended outcome, but where the performance of the action is both primarily and secondarily unreliable. For example, returning to Bishop's example of the fastidious assassin, suppose that the assassin not only plans to kill the tyrant specifically with a shot through the heart, but is also using an unreliable rifle that has a mechanism that tends to jam. In the particular instance in question, the gun operates without jamming, successfully firing the bullet, which then strikes the tyrant's head rather than his heart. In this case the result is not brought about by dependable stages, since the gun has a tendency to jam, and there is also a mismatch with the assassin's plan, yet the killing of the tyrant seems clearly intentional. By stipulating that the firing mechanism is unreliable, a similar example could be constructed using Pears example of the gunman whose off aim is corrected by a crosswind.

In conclusion, neither Bishop's nor Pears' proposals offer an adequate treatment of examples of consequential deviance. Even if their accounts were adequate, they could only offer half of the story since they would not be applicable to the problem of antecedential deviance. As Bishop acknowledges, cases of antecedential deviance, such as Davidson's example of the climber, involve basic actions consisting of directly realizable bodily movements where there is neither the necessity nor the opportunity for the agent to engage in planning about how to realize her goals. Nor does Pears' condition of primary reliability seem to be a necessary condition of intentional basic action. A person recovering from paralysis in her left arm may intentionally move her fingers even though her attempts to do so frequently fails because the physiological

pathway leading from her brain to the muscles in her hand is an unreliable one. In such a case there is primary unreliability in the causal chain. Nonetheless, the agent's basic act of closing her fist is intentional.

I turn now to Bishop's treatment of antecedential deviance. In order to deal with examples in which the deviance affects the chain between the agent's mental state and her initial bodily movement, Bishop appeals to what he refers to as the "sensitivity strategy" (1989: 148). According to the sensitivity strategy, a given episode of behaviour counts as intentional if and only if it displays a certain kind of sensitivity to the content of the intention that was its cause. The behaviour or bodily movement in question in examples of antecedential deviance is not intentional because it fails to demonstrate the requisite kind of sensitivity. Bishop considers different ways of spelling out this sensitivity condition. The first, which Bishop ultimately rejects, involves a counterfactual analysis. According to this proposal, the sensitivity condition is met when there is a relation of counterfactual dependence between the content of the intention and the resulting behaviour such that,

"had the agent's intention differed in content, the resulting behaviour would have differed accordingly." (1989: 150)

This counterfactual analysis of the sensitivity condition stipulates that a given episode of behaviour qualifies as intentional action if and only if had the agent had a slightly different intention then it would not have been the same action that occurred, but one that was appropriately different. In a case such as Davidson's climber, it seems that had the climber's intention been slightly different – say, had she intended to let go of the rope in twenty seconds time – this would not have made a difference to the outcome. She would still have been unnerved by her intention, which would have then caused her to tremble and drop the rope. Therefore, according the counterfactual analysis of sensitivity, she did not let go of the rope intentionally. However, the problem that Bishop finds with the counterfactual analysis is that it is susceptible to the same kind of counterexamples as a parallel hypothesis, defended by David Lewis (1980), for a causal theory of perception. According to Lewis,

"What distinguishes our cases of veridical hallucination from genuine seeing…is that there is no proper counterfactual dependence of visual experience on the scene before the eyes. If the scene had been

different, it would not have caused correspondingly different visual experience to match the different scene" (1980: 245).

As Martin Davies (1983) has argued, it is possible to imagine cases in which Lewis' condition of counterfactual dependence is met, but where it seems that the subject is not having a genuine perceptual experience. Davies (1983: 412) presents an example in which a subject's natural or prosthetic eye is not functioning, but in which a deviant causal chain is operative that causes visual experience in the subject that matches the scene before her eyes. Nonetheless, if the scene were any different the subject's visual experience would be correspondingly different because a 'reverse censor' is standing by who would intervene and permit causation of matching visual experience by the normal means. Bishop says that parallel counterexamples involving a "reverse behavioural censor" (1989: 153) can be constructed against the counterfactual analysis of the sensitivity condition for intentional action. Presumably, Bishop is imagining a case, such as Davidson's climber or Morton's Leo, involving antecedential deviance, but in which if the agent's intention were even very slightly different then an omnipotent and omniscient neurophysiologist – the 'reverse behavioural censor' - would intervene to restore normality to the causal pathway between intention and bodily movement. In such an example it would be true that had the agent's intention differed even minimally then the resulting behaviour would have differed accordingly. However, the bodily movement that actually occurred would still not be an intentional action.

The second way of spelling out the sensitivity condition Bishop considers appeals to Christopher Peacocke's notion of differential explanation. Bishop only very briefly explains the notion of differential explanation. So in order to get a clearer view, it will be helpful to turn to Peacocke's own presentation of the idea in his book, *Holistic Explanation* (1979), where he uses the concept of differential explanation in order to provide an analysis, not just of intentional action, but of genuine perceptual experience as well. Peacocke distinguishes between two types of differential explanation, 'weak' and 'strong'. Peacocke argues that in order for one event to weakly differentially explain a second, there must be a covering law that includes a mathematical function specifying that the one event must follow from the other. Peacocke writes,

"we may offer this rough definition: *x*'s being ϕ [weakly] differentially explains *y*'s being ψ iff *x*'s being ϕ is a non-redundant part of the explanation of *y*'s being ψ, and according to the principles of explanation

(laws) invoked in this explanation, there are functions…specified in these laws such that *y*'s being ψ is fixed by these functions from *x*'s being ϕ." (1979: 66)

In order for one event to *strongly* differentially explain a second, Peacocke says that not only must the first event weakly differentially explain the second event, but the former must be 'stepwise recoverable' from the latter. Peacocke defines stepwise recoverability as follows:

"We have stepwise recoverability of *p* from *q* iff in the explanatory chain from *p* to *q*, at each stage, given just the initial conditions of that stage other than the explanandum of the previous stage, and given also the explanandum of the present stage and its covering law, one can recover the explanandum of the previous stage." (1979: 80)

Stepwise recoverability requires that at each stage in the causal chain between the two events in question it should be possible to deduce the occurrence of any earlier stage in the chain from knowledge of a covering law specifying a function such that the occurrence of the later stage is fixed by the occurrence of the earlier one, along with knowledge of the initial conditions of the later stage (excluding the occurrence of the earlier stage). According to Peacocke, strong differential explanation is necessary and sufficient for sensitivity in the causal chain between two events.

Peacocke applies the notion of differential explanation to the problem of deviant causal chains. He argues that in order for the occurrence of an event or episode of bodily movement to qualify as intentional action it must be the case that the event or bodily movement in question is strongly differentially explained by the agent's possession of an intention with an appropriate matching content. On the basis of the principle that every psychological state is realized by some physical state, Peacocke says that the laws invoked in the differential explanation of intentional action would refer to the agent's intention under its purely neurophysiological description. In cases of causal deviance, Peacocke thinks that the event or bodily movement in question would not be strongly differentially explained by the neurophysiological realization of the agent's intention, but by other factors. For example, in cases of antecedential deviance such as Davidson's climber, in which the agent acts out of nervousness, he thinks that the bodily movement would be strongly differentially explained, not by the agent's intention, by the physiological states with which her state of nervousness interacts to produce bodily movement that just so

happens to match her intention (1979: 70). Why exactly Peacocke thinks that in cases of antecedential deviance the agent's bodily movement would not be strongly differentially explained by her intention is not clear to me. Assuming some suitably formulated version of determinism is true - an assumption that Peacocke would presumably want to grant - and assuming one were a Laplacian demon with complete knowledge of the laws of nature and of the initial conditions of the occurrence of every event, as well as an intellect vast enough to store and compute all this information, then, ignoring the occurrence of quantum indeterminacies, surely *any* event would be stepwise recoverable from any other event that it was in *some way* causally relevant to, no matter how circuitous or distended the causal chain. In which case, it is not clear how exactly one event's being strong differential explanation by another is supposed to exclude deviant causation.

Nonetheless, in Peacocke's view, his analysis of the sensitivity condition in terms of strong differential explanation can cope with a variety of cases of deviant causation. However, he admits that there is one particular kind of example of deviance that it fails to exclude. These are examples of what Bishop refers to in his discussion as 'heteromesial' causal chains (1989: 125), in which the causal chain running from the agent's intention to the intended bodily movements passes through the intentions of another agent. Peacocke presents the case of a knowledgeable neurophysiologist who has blocked the normal causal route between an agent's intentions and her bodily movements. The neurophysiologist has the ability to identify from the agent's neurophysiological states what her intentions are as soon as she forms them. On a particular occasion, the neurophysiologist decides to produce in the agent exactly the motor impulses needed to realize what the neurophysiologist knows to be her intention (1979: 87). According to Peacocke, in this sort of example the agent's bodily movement or behaviour would be strongly differentially explained by her intention. However, Peacocke argues that it is implausible to say that the agent's bodily movement is intentional for the reason that when we say that "an event is, under a given description, intentional of a person, we normally imply that that person was the originator of that event" (1979: 88). In order to exclude this sort of example, Peacocke supplements his analysis of sensitivity with the additional requirement that "the chain from intention to bodily movement should not run through the intentions of another person" (1979: 88).

Though Bishop's own account of intentional basic action, and so of the basis of the distinction between antecedential deviance and nondeviance, incorporates Peacocke's analysis of sensitivity, Bishop argues that supplementing that analysis with a blanket exclusion of examples of heteromesy is not an adequate solution to the problem of heteromesial causal chains. Bishop's reason for claiming this is that we can imagine cases in which the causal chain running from an agent's intentions to her bodily movements do pass through the intentions of another agent, but in which the bodily movements still appear to be intentional actions. To illustrate, Bishop imagines an example in which an agent is fitted with a prosthetic neural replacement that breaks down, but is temporally restored to working order by a second agent who intentionally holds its broken wires together until they can be re-soldered (1989: 159). Bishop argues that in an example of heteronomy of this type the agent with the neural replacement retains direct control over her actions. The actions of the second agent merely form part of a prosthetic aid to action and do not introduce deviant causation. However, he says that in a heteromesial case such as Peacocke's, the agent's behaviour or bodily movement is not intentional because the involvement of the second agent "*preempts or blocks the agent's exercise of direct control*" (1989: 159). According to Bishop, the problem for Peacocke is that his supplementary condition requiring that the chain from intention to bodily movement should not run through the intentions of another person excludes both types of cases of heteromesy when it is only the latter, preemptive cases that are genuinely problematic.

In developing his own account of the basis of the distinction between deviance and nondeviance in cases of antecedential deviant causal chains, Bishop sets himself the task of identifying a further necessary condition beyond sensitivity that, rather than constituting a blanket exclusion, can accommodate the heterogeneity of examples of heteromesy. He appeals to an idea defended by Irving Thalberg that intentional basic action involves "sustained causation". Bishop writes that for an episode of behaviour to be causally sustained is for it to be "continuously regulated" by the agent (1989: 167). He argues that in cases of preemptive heteromesy it is precisely this sustained causation, or continuous regulation, by the agent that is missing. His basic idea seems to be that the relevant difference between the preemptive heteromesial case and the case of the heteromesial prosthetic aid is that in the former it is the neurophysiologist who is controlling, or, in Bishop's terms, "continuously regulating", the movement of the agent's body, whereas in the latter the second agent is merely *enabling* the first agent to

persist in controlling, or continuously regulating, the movement of her own body. However, Bishop notes that it would not be enough to simply supplement the sensitivity requirement on intentional action with a requirement that there be sustained causation, or continuous regulation, by the agent since sustained causation "is a matter of *the agents' doing something*" (1989: 167). Thus what Bishop needs is a non-circular way of unpacking the notion of sustained causation or continuous regulation in strictly event-causal terms. He appeals to a notion, borrowed from cybernetics, of a 'servosytem'. A servosystem is a mechanism of feedback that is incorporated within a system of control, such as a thermostat, and that produces and maintains an output by means of such feedback. Bishop maintains that in order for a bodily movement to qualify as the intentional action of an agent, the feedback involved in the servosystem linking intention with behaviour must be routed via the agent's *own* central processing system, where Bishop is referring specifically to the processing of "feedback information about orientation and muscular states" (1989: 170). So in the case of Peacocke's knowledgeable neurophysiologist, the reason why Bishop thinks that this counts as an example of deviance is that the feedback mechanism through which the agent's behaviour is continuously regulated is not routed through the agent's own central processing system, but through the central processing system of the neurophysiologist's computer. Bishop acknowledges that not all cases of intentional action do seem to involve sustained causation or controlled regulation involving feedback via the agent's own central processing system – consider, for example, a person intentionally blinking her eye. For this reason, he only defends the weaker, conditional requirement that *if* the causal pathway from intention to resultant bodily movement does involve feedback loops, then it must be to the central mental processes of the agent that the feedback returns, as opposed, say, to those of some intervening neurophysiologist (1989: 171). To summarize, Bishop holds that an agent "*M* performs the basic intentional action of *a*-ing if and only if,

> *M* has a (basic) intention to do *a*; and,
>
> > *M*'s having this basic intention causes *M* to produce behaviour, *b*, which instantiates the types of state or event intrinsic to the action of *a*-ing; where
> >
> > (i) the causal mechanism from *M*'s basic intention to *b* satisfies the sensitivity condition [where Bishop appears to want to analyze this condition in in accordance with Peacocke][6]; and

[6] Though Bishop does not explicitly endorse Peacocke's analysis, given his objections to the counterfactual analysis I interpret him as understanding sensitivity in terms of differential explanation.

(ii) if this causal mechanism involves feedback, then the feedback signal is routed back to *M*'s central mental processes if to anyone's." (1989: 172)

It seems to me that there are three main problems with Bishop's account. Firstly, for the reasons stated above, it is not clear to me that the analysis of sensitivity either in terms of counterfactual dependence or in terms of differential explanation is sufficient to exclude examples of antecedential deviance. The counterfactual analysis is not sufficient because, as Bishop himself points out, we can imagine cases of antecedential deviance in which had the agent's intention been even slightly different then the appropriate matching behaviour would have occurred. This is the point of Bishop's 'reverse behavioural censor'. On the other hand, Peacocke's analysis seems insufficient because it is not at all clear why he thinks that in cases of antecedential deviance the agent's bodily movement would not be strongly differentially explained by the neurophysiological realization of her intention. As I suggested above, assuming some suitably formulated version of determinism, and assuming one were a Lapacian demon, then, ignoring quantum indeterminacies, any event could be strongly differentially explained by any other event that it was in some way causally related to.

The second problem is that, given that Bishop only defends the conditional claim stated in (ii), he has provided no necessary condition beyond sensitivity for intentional basic action that does not involve sustained causation or controlled regulation – that is, for intentional action where the causal mechanism from intention to bodily movement does not involve feedback. As already mentioned, Bishop acknowledges that there are many kinds of intentional basic actions that do not involve sustained causation by the agent. Given that counterexamples to the sufficiency of the sensitivity condition work just as well for these kinds of intentional actions – imagine that Peacocke's knowledgeable neurophysiologist controls the agent's body in conformity with her intention to blink her left eye - this is a major lacuna in Bishop's account. Thus the second problem with Bishop's account is that it is incomplete.

A final objection, which has been pressed by David Hillel-Ruben (1991), relates to the necessity of Bishop's condition (ii). Hillel-Ruben presents an example in which it would appear that (ii) is not met, but where the agent's actions still look to be intentional. The example is a variation on Bishop's case of the heteromesial prosthetic aid. In Hillel-Ruben's version, the agent's damaged nerves have been replaced by prosthetic ones.

These are broken and are temporarily held together by a second agent, thereby enabling the first agent to act on her intentions. However, in so doing - and this is where Hillel-Ruben's example departs from Bishop's - the second agent has to take account of a correlation between how the first agent moves and how the prosthetic nerve endings must be fitted together. As the first agent moves, the second agent must observe and take account of these movements and adjust the manner in which the nerve endings are connected (1991: 288). As it stands, this story may not seem entirely coherent. For if the second agent observes the first agent's movements before adjusting the wires accordingly then it would seem that the first agent has already moved as she intended, and therefore does not require the assistance of the second agent. To remove any potential incoherence here, it will help to fill in a few details. Suppose that the first agent's prosthetic aid is only partially faulty and that she is able to complete a certain stage of the action she intends to perform on her own, but requires the assistance of the second agent for the completion of the rest. For example, suppose that the first agent intends to type a word on her laptop. She is able to move her finger to the location of the relevant key unassisted. Her assistant must then note the specific finger that she wishes to use, observe when it is directly above the relevant key, and then adjust the prosthetic nerve endings in such a way that the first agent is able to intentionally press down on the relevant key with the relevant finger and then release it. In an example such as this, the second agent has become an essential constituent of the causal pathway between the first agent's intention and her bodily movement. She is directly involved in the processing of feedback information about orientation and muscular states on which the completion of the action depends. Nonetheless, the bodily movement of the first agent is intentional. It is the first agent who controls her own movements in conformity with her intention and types the word that she has in mind using the temporally restored prosthetic aid. The actions of the second agent merely serve as enabling her to do so. One thing that Bishop might say here could be that Hillel-Ruben has actually presented us with a case of joint or shared intentional action. The typing of the sentence cannot properly, or without qualification, be said to be an intentional action of the first agent because it is an intentional action of both collectively. Therefore, this is not a counterexample to his account. Adequately addressing this response would require straying into the metaphysics of group agency and extended mind and would take me too far afield for the purposes of this chapter. However, I hope it will be sufficient to note two things. Firstly, there is a strong intuition of clear and distinct individual agency in

the example. There is a strong intuition that it is the first agent who intentionally types the sentence, just as she would with an undamaged prosthetic aid, and that it is the second agent who intentionally holds and adjusts the nerve endings in the appropriate manner, restoring the prosthetic aid to working order. Secondly, if Bishop was willing to attribute shared agency to the two agents in the Hillel-Ruben example then, ostensibly at least, it seems that he should also attribute shared agency in Peacocke's example of the knowledgeable neurophysiologist since both involve heteromesy. However, Bishop takes the latter example to be a clear instance of intentional action by a single agent. He views the movements of the first agent's body as intentional actions of the neurophysiologist. I suggest that consistency demands the same judgment in the Hillel-Ruben case.

I conclude that in the end neither Bishop's account of the distinction between normal and wayward causation in examples of consequential deviance, nor in examples of antecedential deviance, is adequate. Given that Bishop's event-causal response to the problem of deviant causal chains is the most developed that I am aware of, and given the problems with his account that I have outlined, I believe that we have good grounds for seeking a non-event causal solution to the problem. It is to this task that I now turn.

5.2 Hyman's Solution to the Problem of Causal Deviance

I turn now to John Hyman's argument. In a chapter of his forthcoming book, *Action, Knowledge and Will*, Hyman defends the view that all intentional action is caused by desire. He argues that the possibility of deviant causal chains is only problematic for this view when we conceive of desires as causes on the model of the Humean theory of causation. By the 'Humean theory', Hyman is not referring necessarily to Hume's own theory of causation, of which he does not offer an exegesis, but "to various combinations of doctrines or assumptions about causation which were defended by or have been imputed to or derive from Hume" (*forthcoming*: 7). Hyman argues that once we analyze the causal relation between desire and intentional action, not along Humean lines, but rather as the exercise of a disposition the possibility of deviant causal chains ceases to be problematic[7]. I begin this section by explaining Hyman's dispositional account of desires. I will then

[7] Substantially the same line of argument can be found in Hyman, 'Desires, Dispositions and Deviant Causal Chains', forthcoming in *Philosophy* 89 (January 2014).

explain why Hyman thinks that understanding the causal relation between desire and intentional action in Humean terms gives rise to the problem of causal deviancy. Finally, I will explain how Hyman thinks that his dispositional analysis of desire avoids the problem.

According to Hyman, "intentional action always (or almost always) involves desire" (*forthcoming*: 2). By 'desire', Hyman says that he means this in a "broad sense…[that] refers to the whole gamut of wanting and willing" - a sense in which he says we might also refer to it as "will" (*forthcoming*: 3). He contrasts his own analysis of desire with that of Bertrand Russell in Russell's, *The Analysis of Mind*. Hyman construes Russell's account of desire in terms of three claims:

> "First, 'desire must be capable of being exhibited in action'; second, the action that exhibits a desire is caused by a 'mental occurrence involving discomfort'… such as a disagreeable sensation; third, an animal is said to 'desire' the action's 'purpose', which Russell defines as 'the result which brings it to an end', 'the state that brings quiescence'" (*forthcoming*: 3).

Hyman agrees with the first of these claims. However, he rejects the second and third. The second claim states that action exhibiting desire is always caused by a mental occurrence involving discomfort. Hyman concedes that this might be true of actions motivated by urges and cravings, but that it is not generally true of action involving desire. Rather, he claims that some purposive behaviour is motivated by a desire to experience something pleasant, or, more generally, the desire to achieve some goal or realize some aim. Hyman's objection to the third claim – the claim that the object of desire is the thing that brings quiescence – appeals to a criticism made by Wittgenstein. Wittgenstein noted that a punch in the stomach may make the desire to eat go away. However, a punch in the stomach was not what the desire to eat was for. Hyman views this objection as decisive.

While Hyman rejects Russell's theory, he thinks that it contains two plausible ideas. One is the idea, already mentioned, that "desire must be capable of being exhibited in action" (*forthcoming*: 4). In other words, for a desire to or for x to be attributable to an agent there must be some kind of behaviour that would count as a manifestation of that desire. Thus Hyman writes that "There is such a thing as wanting to die, but there is no such thing as wanting (as opposed to wishing) one had not been born" (*forthcoming*: 4). The

second aspect of Russell's account that Hyman finds plausible is the idea that "desire is intrinsically related to pleasure and pain" (*forthcoming*: 4). Hyman explains this as meaning that it is necessarily the case that satisfying a desire is something pleasant whereas failing to satisfy a desire is something unpleasant. Hyman claims that these two ideas – that desire is intrinsically related to pleasure and pain, and that desire must be capable of being exhibited in action – support the view that desires are dispositions. However, Hyman thinks that while desires are dispositions, they are not dispositions like ordinary physical dispositions. He argues that whereas a simple physical disposition, such as solubility, is manifested in causing or undergoing some sort of change, "a desire is manifested by both action and reaction" (*forthcoming*: 4). Hyman says that the 'action' that manifests or expresses desire is whatever is done with the aim or intention of getting or doing what the desire is for (*forthcoming*: 4). He says that the 'reaction' is "feeling glad, pleased or relieved if the desire is satisfied and sorry, displeased or disappointed if it is frustrated" (*forthcoming*: 5).

Hyman argues that the active manifestation of desire, unlike the manifestation of simple physical dispositions, is teleological or goal-directed. He thinks that desires are essentially dispositions to pursue an aim, goal or end. Hyman regards this as a difference between the manifestation of desire and the manifestation of simple physical dispositions such as "the disposition of a floating body to displace liquid" (*forthcoming*: 5). However, Hyman does not think that this alone distinguishes the active manifestation of desire from the manifestation of the dispositions of more primitive or simpler life forms, such as the phototropism of plants, which he thinks is manifested or expressed in teleological or goal-directed behaviour. He cites two further reasons for distinguishing desires from simple physical dispositions, and that also distinguish them "even from primitive behavioural dispositions like the phototropic tendencies of plants" (*forthcoming*: 5). Firstly, he says that the object of desire and its manifestation need not be identical, but can be related as end and means. Second, Hyman notes that the way in which desire is expressed in action depends on cognitive factors – on what the agent knows and believes. In other words, individuals might express or manifest the same desires in different ways depending on their cognitive backgrounds of knowledge or beliefs about the world.

Hyman anticipates the worry that if desires are dispositions then desires must explain action in the way that dispositions in general explain phenomena. However, it is sometimes said that explanations appealing to dispositions are vacuous or uninformative. The classic example is Moliere's joke, which involves the student who explains why opium makes one fall asleep by appeal to it 'virtus dormitiva' – i.e. its soporific power or disposition. Hyman responds, legitimately it seems to me, that even in the case of Moliere's joke, the explanation is not entirely uninformative. He argues that while the explanation does not tell us very much, it does tell us something – namely, that the cause of falling asleep after ingesting opium is not something to do with the interaction of the opium with some other agent ingested by the subject, but is intrinsic to the opium itself (*forthcoming*: 6). While Hyman acknowledges that in an example such as this the appeal to a disposition is not hugely informative, there are many other examples of explanations appealing to dispositions that are considerably more informative. He gives the example, 'Lead is poisonous because it is a neurotoxin', which he says is more informative than the example from Moliere because it "excludes a larger range of alternatives". Similarly, he argues that, 'the agent ϕ'd because he wanted to', may not be very informative, but, 'he opened the window because he wanted to hear what the neighbors are saying', is (*forthcoming*: 6). Once again, this is because it excludes a larger range of alternatives.

Hyman argues that the reason why theorists, such as Davidson, who defend a causal conception of intentional action, have been unable to resolve the problem of deviant causal chains is that they cleave to what Hyman calls the 'Humean theory' of causation. Hyman does not offer any systematic exposition of precisely what he means by the 'Humean theory'. Rather, he associates it with the following assortment of claims:

"Causes are events, which are distinct from and precede their effects…

Our knowledge of causes is uncertain, and relies on inductive inference from a plurality of cases…

A singular causal statement is implicitly general: it implies that events resembling the cause are invariably followed by events resembling the effect…

It is conceivable that any kind of cause should have had a different kind of effect from the one it actually has…" (*forthcoming*: 7)

For Hyman's purposes, I think that the most important of the above are the first and the last claims. According to the first claim, on the Humean theory the relata of the causal relation are always events. This automatically excludes dispositions from being causes, which are not event, but properties. The last claim relates to the fact that Hume denies the existence of any necessary connection between causes and effects. Hume argues that the idea of there being a necessary connection between events derives from the feeling of habitual expectation that arises from having observed a regularity or constant conjunction between occurrences. However, for Hume, the idea of necessary connection is essentially confused. On Hume's account, there are no causal powers or dispositions intrinsic to objects that constitute the metaphysical glue binding one event to another, and in virtue of which one event can be said to have necessitated a second. Rather, the fact that one event led to another is always contingent. When Hyman speaks of the Humean theory of causation, he is referring to a tradition in the philosophy of causation that has focused on trying to spell out how we should understand the causal relation without appeal to the notion of necessary connection.

Hyman argues that the Humean theory of causation inevitably invites the objection from causal deviance. He writes:

> "if causation consists simply in the regular or law-like concatenation of events, one causal chain may be more circuitous than another, but the question 'Deviant or normal?' cannot arise, because there is no power in the world from which an action or event can either proceed, if the chain is normal, or fail to proceed, if it is not. This is why the Humean theory of causation cannot 'eliminate the deviant causal chain', for example, why it cannot distinguish between the case where the climber loosens his hold *in order to* kill the other man and the case where his desire to kill the other man causes him to loosen his hold without him doing so for this reason. This is just a particular instance of a more general failure to distinguish between 'true' causes and 'occasional' causes." (*forthcoming*: 14).

According to Hyman, on the Humean conception of causation, a causal process consists in a regular or law-like concatenation of events. However, it is always possible to think up a way of supplementing or substituting one or more of the links or stages that we standardly expect to find in the causal sequence or chain with one or more links or stages that we would not normally expect to find. In Hyman's words, it is always possible to think up a chain or sequence that is more 'circuitous'. Thus for any 'normal' chain of events - *A*, *B*, *C* – we can always think of an abnormal or deviant chain – *A*, *B*, *B**, *C*. We want to be able to say that in the case of the latter, abnormal chain, unlike the

former, normal one, *A* is *a* cause of *C*, in the sense of being causally implicated in the occurrence of *C*, but that it is not *the* cause, or, as Hyman puts it, it is not the 'true' cause. However, Hyman thinks that the Humean theory gives us no adequate basis for distinguishing between the two cases in this way. I think that another way of making the same point would be to say that the Humean theory gives us no account of the distinction between the cause of some event and a mere condition of that event. In the deviant chain, *A* is not the cause of *C*, but a mere condition of *C*. However, Hyman's claim is that the Humean theory does not have the resources to adequately explain what this difference amounts to. We can illustrate this with reference to the heteromesial counterexamples to Peacocke's and Bishop's accounts, where the abnormal or deviant link in the chain involves, not, for instance, the state of nervousness, as in Davidson's example the climber, but the interventions of a second agent. In the heteromesial cases, even though Peacocke's sensitivity requirement is met, the agent's intention is still not the 'true' cause of her bodily movement. It is merely a condition of it. The difficulty for Peacocke and Bishop is to explain why exactly this is, and hence why exactly the chain counts as deviant.

Hyman argues that the problem with the Humean theory is that it does not make any reference to the existence of powers or dispositions. He thinks that it is because the Humean theory does away with powers that it cannot distinguish 'true' causes from 'occasional' causes. According to Hyman, it is not only the case that some events are causes of other events, but it is also the case that some events are the manifestations of powers or dispositions, where the latter constitutes a distinctive and important kind of causation that is not merely reducible to the former, event kind. Once we introduce powers and dispositions into our ontology, we can say that some power or disposition, *x*, is the 'true' cause of some event, *e*, when *e* is a manifestation of *x*. And we can say that *x* is merely *a* cause of *e*, as in examples of deviant or wayward causal chains, when *e* is not a manifestation of *x*, but when *x* is nonetheless in some way causally implicated in bringing about *e*. According to Hyman, in Davidson's example of the climber, the climber loosening his hold on the rope in not intentional because it is not a manifestation of the climber's desire to rid himself of the danger. In this sense, the climber's desire is not what Hyman calls the 'true' cause of his letting go of the rope. Nonetheless, it is *a* cause since it led to his state of nervousness, which in turn resulted in his letting go. Likewise, I want to suggest that in Peacocke's example of the knowledgeable neurophysiologist,

the agent's bodily movement is not intentional because it is not a manifestation of her *intention* to so move her body. In this sense, her intention is not the true cause of her bodily movement. Nonetheless, it is *a* cause since it led to the actions of the neurophysiologist, whose intention to cause her, the first agent's, body to move in conformity with her, the first agent's, intentions was the true cause of the bodily movement.

Hyman resists the idea that talk of dispositions and their manifestation can be reduced to the language of event causation. Hyman notes that many contemporary philosophers have attempted to offer some sort of reductive analysis of dispositions that clarifies or reforms our talk of them. He considers what he refers to as "the 'simple' or 'naïve' conditional analyses":

"(C) A substance x is disposed at time *t* to give response *r* to stimulus s if, and only if, x would give response r if x were to receive stimulus s at time t" (*forthcoming*: 14).

However, Hyman is skeptical about the prospects of any such reductive analysis. He argues that the conditional analysis of dispositions is open to various sorts of counterexamples. He describes two main kinds. The first kind of counterexample to the conditional analysis involves cases of 'finkish' dispositions. Hyman presents a famous example from C.B. Martin (1994). A wire is live if and only if it is disposed to transmit a current when touched by an earthed conductor, and it is dead if and only if it is not live. In Martin's example, a live wire is attached to a safety-device that detects when the wire is touched and makes it dead before the current begins to flow. Therefore, since the wire is live it is disposed to transmit a current when touched, but nonetheless it will not transmit a current. So the left hand side of the bi-conditional, (C), is true, but the right hand side if false. The second kind of counterexample to the conditional analysis that Hyman describes involves cases of 'masked' dispositions. Hyman presents the example of a sheet of toughened glass fitted with an explosive devise that is sensitive to being dropped. In this case, the glass is not disposed to shatter when dropped because it is toughened, but nonetheless it would shatter when dropped due to the explosive device. Therefore, the left hand side of the bi-conditional is false, but the right hand side is true. Hyman observes that in both sorts of cases there is an abnormality in the situation that is due to some factor that is extrinsic to the object with the disposition in question. Thus one might attempt to defend the conditional analysis of dispositions by supplementing it

with some sort of ceteris paribus clause requiring that the situation be normal or 'non-deviant' in the sense of excluding such extrinsic influences. However, Hyman argues that any reductive analysis of dispositions that aims to clarify or reform our talk of them will need to be specific. It will have to say precisely what sort of process or situation counts as normal, and so what sorts of extrinsic factors should be excluded. However, this will be extremely complex and will differ for each particular kind of disposition. Hyman argues that if we cannot specify precisely when other things are equal for each particular disposition then this means that all we can really say is that other things are equal, or the process is normal, if and only if the disposition in question is manifested (*forthcoming*: 17). However, if this is all we can say then it seems the conditional analysis has failed in what it set out to do.

How is Hyman's dispositional analysis of desire meant to solve the problem of deviant causal chains? I think that the heart of Hyman's solution to the problem of deviant causal chains lies in the fact that, to borrow Alexander Bird's phrase (2010: 162), dispositions can be 'mimicked'. Hyman says that almost every disposition can be connected to the kind of occurrence that would ordinarily count as its manifestation via some sort of deviant causal chain (*forthcoming*: 10). Hyman gives the example of a man who takes a soporific before driving. The drowsiness induced by the drug causes the man to have an accident in which he is knocked unconscious. Hyman argues that if this were to happen then the ingestion of the drug would have been a cause of the driver's losing consciousness. However, the driver's losing consciousness would not have been a manifestation of the soporific disposition of the drug (*forthcoming*: 10). According to Hyman, there is no purely event-causal analysis that can explain the sense in which this is a deviant causal chain. Rather, all we can say is that the chain was deviant because the disposition of the drug was not manifested. Its manifestation was preempted by the accident. Hyman's example is an instance of the more general phenomenon that Bird refers to as 'mimicking'. The mimicking of a disposition is when an event of the same type as the manifestation of a particular kind of disposition occurs, yet does not seem to count as a manifestation of that disposition. We might consider a different example. Suppose that a sheet of glass is disposed to shatter when dropped. The sheet of glass has an explosive device attached to it. If the glass were dropped then the device would be set off within the first nanosecond of its achieving contact with the ground. This would result in the shattering of the glass. However, the shattering of the glass would not be a

manifestation of its disposition to shatter when dropped. The manifestation of that disposition would have been preempted by the setting off of the explosive device. Once again, there is no purely event-causal analysis that can explain why this is a deviant causal chain. Certainly the event would be unusual relative to how sheets of glass normally shatter on dropping. However, even in a world in which such glass-shattering events were the norm there would still be a clear sense in which it involves deviance. The sense in which it involves deviance cannot be explained without reference to the disposition. The chain is deviant because the disposition was not manifested. Similarly, in Davidson's example of the climber, or in the examples of heteromesial deviance, Hyman's view is that all we can say is that the chain was deviant because the agent's intention was a cause of, but was not manifested in, the bodily movement that eventuated. I suggest that Hyman's underlying idea is that we think of these two sorts of cases, that of the glass shattering as a result of the explosive device, and that of the examples of deviance typically invoked in the literature against the causal theory of intentional action, as exactly parallel. In both sorts of cases we have the occurrence of an event with precisely the same intrinsic physical properties as the type of event that would count as a manifestation of the disposition, yet the disposition is not manifest. In both cases there is no event-causal analysis that can explain why the disposition in question is not manifest. Rather, the relation between a disposition and its manifestation is primitive and not susceptible to event-causal analysis. According to Hyman, once we think of desires as dispositions typical examples of causal deviance such as these cease to look problematic. They are simply further examples of the mimicking of dispositions.

Hyman's account of intentional action seems to involve an interesting and unusual combination of claims that have generally been treated as contrary to one another. On the one hand, Hyman defends a causal account of intentional action. For Hyman, an event or episode of behaviour is intentional if and only if it is caused in the right sort of way by a desire with an appropriate content. On the other hand, Hyman appears to view intentional action as something that is irreducible and conceptually primitive. It is not amenable to event-causal analysis. According to Hyman, an event or episode of behaviour is intentional if and only if it is a manifestation of desire. With regards to precisely what constitutes a manifestation of desire Hyman seems to be of the view that there is little more to be said. On the whole I find Hyman's response to the problem of deviant causal chains extremely convincing. My main objection is to Hyman's

characterization of desire. As previously mentioned, Hyman states that he is analyzing 'desire' in its "broad sense" and in a way that could also be referred to as "will". It seems to me that what Hyman means by 'willing' is simply intending, and that Hyman's dispositional analysis of desire is in fact an analysis of the attitude of intention. In my view, it is intentions that are dispositions made manifest in goal-directed behaviour, conceived broadly so as to include both overt action, as well as the mental activities of planning and reasoning emphasized by Bratman. I do agree with Hyman that everything that a person does intentionally she can be said to 'want' to do in some broad sense of the term. But the sense of 'wanting' here is that of 'preference'. Whatever an agent does intentionally she chooses or prefers to do. And, for reasons presented by Bratman and discussed in chapter one, what a person chooses to do, what Bratman calls the 'conclusion of her practical reasoning', is not the same as what she intends. What a person chooses or prefers is some overall state of affairs. On the other hand, the object of an intention is always some end or some means to an end. It is some achievement or activity. It is not necessary for me to argue that there are no actions that are manifestations of desire. Perhaps certain crimes of passion, or actions performed in a state of temporary insanity, are manifestations of desire. Perhaps there are also certain Freudian cases involving actions that manifest desire, rather than intention. At one point, Velleman (2000: 2) considers an example drawn from Freud in which Freud's sister tells Freud that his new writing table looks attractive, except that his inkstand does not match. Later, after his sister has left the room, Freud sweeps the inkstand off the table with an apparently clumsy movement, smashing it. On Freud's interpretation, his movement was not clumsy at all, but was in fact a well-executed action motivated by his desire to break the inkstand and his belief that this would result in his sister buying him a new one. If Freud's analysis is correct then perhaps this is a case of action that manifests desire. However, this is not an instance of intentional action. This is because it is not an action that manifests intention.

The dispositional analysis of intention has a number of virtues. It explains how an intention may not only be revealed or exhibited *in* action, but may be future-directed. For dispositions may exist latently without being currently manifested, even without being manifested at all. Thus it explains how there can even be what Davidson refers to as 'pure intending' (1980: 83) – i.e. having an intention that is never in any way acted upon. This proposal is also consistent with Bratman's claim that intentions are

associated with propensities not only to act, but also to plan and reason. For planning and reasoning may also be conceived of as manifestations of goal-directed behaviour broadly construed. Thus the analysis of intention as a disposition to pursue an aim or goal can explain how forming intentions, in conjunction with certain other kinds of cognitive capacities, could subserve planning agency, without being metaphysically bound up with planning agency. There could be beings who manifest intention but who are not planners. Finally, I have argued that a proper appreciation of the fact that intentions are particular kinds of disposition can explain the relation between intention and intentional action. Once we recognize that intentions are a kind of disposition, the possibility of deviant causal chains no longer looks problematic for the causal theory of intentional action.

5.3 Are Dispositions Causally Relevant?

In this section I defend a crucial claim that Hyman makes, but does not defend or expand on, that dispositions are causally relevant to what they manifest. I begin by presenting a prima facie case for thinking that dispositions can be causes. I then attempt to see off a positive argument for the causal irrelevance of dispositions.

According to Hyman, intentional actions are caused by desires, which are dispositions that are manifested in intentional action. Thus Hyman thinks that dispositions are causally relevant to what they manifest. Hyman's reason for thinking this is simply that dispositions have explanatory power. The underlying principle seems to be something like the following:

> *(E)*: A property *P* is causally relevant to some occurrence *O* if *P* could appropriately figure in an explanation of *O*.

Hyman's argument is that dispositions can appropriately figure in explanations of what they manifest. "It's a neurotoxin", would, at least in many contexts, be an appropriate answer to the question, "Why did ingesting lead poison her?" just as, "It's live", would, in many contexts, be an appropriate answer to the question, "Why did touching the wire give her a shock?" In which case, given *(E)*, it would seem that dispositions are causally relevant to their manifestations. I think that this prima facie case can be bolstered when

we reflect on the fact that the concept of causation is bound up and interconnected with a range of other concepts besides explanation. As Jonathan Schaffer writes,

"part of the point of analyzing causation is to shed light on such connected concepts as *prediction*, *explanation*, *manipulability*, and *responsibility*. To a crude approximation, causes license predictions of their effects, causes serve to explain their effects, causes serve as means to manipulate their effects, and causes bestow moral responsibility on agents for their foreseeable effects. While these principles no doubt need refinement, I submit that any event that satisfies every one of these principles (relative to a given effect) deserves to be considered cause-worthy – that event plays all the roles causes plays." (2003: 29)

What Schaffer says of events can also be said of dispositions. If dispositions satisfy every one of these principles then, at least ostensibly, they deserve to be considered 'cause-worthy'. It seems like dispositions do play all of these roles. The fact that x is disposed to ϕ can license a prediction that, other things being equal, under certain conditions, x will ϕ. With regards to manipulability, altering the disposition of x to ϕ in certain ways will alter x's ϕ-ing in corresponding ways. Thus if one toughens a sheet of glass to degree n this will correspondingly alter the effect of dropping it; if one increases the conductivity of some material to degree n by altering its temperature this will correspondingly alter the conduction of an electrical current passing through that material; and if one further unnerves an already nervous waiter this will correspondingly effect the likelihood of her trembling and dropping her tray. Finally, occurrences that are manifestations of an agent's will are evidently occurrences that the agent is responsible for. According to Schaffer, causes bestow moral responsibility on agents for their effects. I claim that it is the disposition to pursue the aim or goal of ϕ-ing that bestows moral responsibility on the agent for ϕ-ing. In which case, prima facie, dispositions are causes.

Nonetheless, even if there is a prima facie case for supposing that dispositions can be causes, it might be argued that there are reasons for thinking otherwise. In the remainder of this section I discuss an argument for the causal impotence of dispositions made by Prior, Pargetter and Jackson (1982). Their argument can be set out as follows:

1. Every disposition has a 'causal basis'. This is to say that for every disposition, any object bearing that disposition has a property or property-complex that, together with the relevant event that constitutes the trigger or stimulus of the

disposition, constitutes the causally operative sufficient condition for the manifestation of the disposition (or, in the case of probabilistic dispositions, the causally operative sufficient condition for the chance of manifestation) (1982: 251).

2. If a property *P* is causally sufficient for the occurrence of some event then no other property distinct from or non-identical with *P* can be causally relevant to the occurrence of that event. (1982: 255).
3. From (1) and (2), no property distinct from or non-identical with the causal basis of a disposition can be an operative sufficient condition of the manifestation of that disposition.
4. Dispositions are properties that are distinct from or non-identical with their causal bases (1982: 253).
5. Therefore, from (3) and (4), dispositions are causally irrelevant to their manifestations (1982: 255).

I will not take issue with premises (1) and (4). My objection to the argument is to (2). Contrary to Prior, Pargetter and Jackson, Schaffer (2003) argues that causal overdetermination involving two distinct operative sufficient conditions is both possible and widespread. Schaffer gives the example of two rocks thrown by vandals that simultaneously shatter a window (2003: 23). He argues that in such a case it is obvious that both the throwing of one of the rocks and the throwing of the other each have something to do with the breaking of the window. However, he considers the question of precisely how we should conceive of the causal relation of these two events to the breaking of the window. On one view, which he calls "individualism" (2003: 24), Schaffer says that the two events of throwing each of the rocks individually cause the breaking of the window. On a second view, which he calls "collectivism" (2003: 24), Schaffer says that the two events of throwing each of the rocks collectively cause the breaking of the window. Schaffer defends individualism. In other words, he defends the view that each rock-throwing event constitutes an operative sufficient condition of the window breaking. Schaffer also considers what he refers to as 'quantitative overdetermination' (2003: 28). According to Schaffer, "Quantitative overdetermination occurs whenever the cause of an event is decomposable into distinct and independently sufficient parts" (2003: 28). He gives the example of a large rock thrown at a window, where the rock's eastern and western hemispheres are each individual overdeterminers of

the shattering of the window (2003: 28). Schaffer provides several arguments in favour of the individualist intuition in these examples. One of these is that "individual overdeterminers play the [aforementioned] predictive, explanatory, manipulative, and moral roles of causes" (2003: 29). However, perhaps Schaffer's strongest argument is that the collectivist can provide no stable account of the causal role of individual determiners – for example, what the throwing of each rock has to do with the smashing of the window. For if the collectivist argues that each individual rock in the two-rocks example, or each hemisphere of the rock in the quantitative overdetermination example, contribute nothing causally to the breaking of the window, then it is difficult to see how the summation of these things could contribute anything. Rather, the pair of rocks in the former example, and the mereological sum of the parts in the latter, seems to possess some "mysterious emergent power" (2003: 38). On the other hand, if the collectivist argues that the individual overdeterminers do contribute something causally to the breaking of the window then it becomes hard to see how it could be denied that each overdeterminer should be counted as a cause. After all, the contribution of each is so significant that by itself it would be sufficient to produce the breaking of the window (2003: 38).

Irrespective of Schaffer's arguments for individualism in the above sorts of cases, I think that Prior, Pargetter and Jackson might argue that such examples do not provide a good model for thinking about the relation between dispositions and their causal bases. Overdetermination by dispositions and their causal bases would involve overdetermination at the level of a higher-order property of an object and at the level of the physical structure on which that property supervenes, whereas the two-rocks case and quantitative overdetermination involve overdetermination by two distinct objects and by distinct parts of a single object respectively. For this reason, they might object that, even if Schaffer' examples establish some sort of precedent for causal overdetermination, examples of the latter kind of overdetermination do not illustrate or make it intelligible how there could be overdetermination of the former kind. However, I think that there other examples of overdetermination presented by Stephen Yablo (1992) that do provide a plausible framework for thinking about overdetermination by dispositions and their causal bases. Yablo's primary concern in presenting these examples is to fend off a parallel argument to the one presented by Prior, Pargetter and Jackson, but for the causal inefficacy of the mental. Yablo refers to this argument as

"the *exclusion argument* for epiphenomenalism" (1992: 246/7), which he summarizes as follows:

"How can mental phenomena affect what happens physically? Every physical outcome is causally assured already by preexisting physical circumstances; its mental antecedents are therefore left with nothing to contribute" (1992: 246)

Yablo notes that the exclusion argument for epiphenomenalism can be posed either as an argument for the causal irrelevance of mental events, or as an argument for the causal irrelevance of mental properties. The crucial premise of the argument as it applies to mental events is as follows:

"If an event *x* is causally sufficient for an event *y*, then no event *x** distinct from *x* is causally relevant to *y* (*exclusion*)" (1992: 247).

Yablo writes that for the version of the argument that applies to properties, 'event x' should be replaced with 'property X' (1992: 247). Applied to the exclusion principle, this gives us,

> *(Ex*)*: If a property *X* is causally sufficient for an event *y*, then no property *X** distinct from *X* is causally relevant to *y*.

(*Ex**) is equivalent to the second premise of the argument attributed to Prior, Pargetter and Jackson above. Yablo rejects the exclusion principle in both its forms, so both as applied to events and as applied to properties. Since I am concerned with defending the causal relevance of dispositions, and since dispositions are properties of their bearers and not events, I will focus on Yablo's rejection of (*Ex**).

Yablo's reason for rejecting (*Ex**) is that it looks false with respect to determinates and their determinables. Yablo gives the following definition of the determination relation:

"P determines Q iff: for a thing to be P is for it to be Q, not *simpliciter*, but in a specific way" (1992: 252).

For example, the property of redness determines the property of being coloured because to be red is to be coloured, not simpliciter, but in a specific way. Equally, the property of

being scarlet determines the property of redness because to be scarlet is to be red, not simplicter, but in a specific way. Yablo offers the following necessary condition of the determination relation:

> "*P* determines *Q* (*P* > *Q*) only if:
>
> (i) necessarily, for all *x*, if *x* has *P* then *x* has *Q*; and
>
> (ii) possibly, for some *x*, *x* has *Q* but lacks p." (1992: 252)

Yablo refers to the relation described by (i) and (ii) as the 'asymmetric necessitation' of *Q* by *P*. According to clause (ii), for any given determinate-determinable pair, *P* and *Q*, and for any *x*, *x* may possess the determinable *Q* without possessing the determinate *P* (1992: 252). Yablo argues that it is a necessary condition of the identity of two properties that it is metaphysically impossible for something to possess one of those properties without possessing the other (1992: 251). It follows that *P* and *Q* are non-identical. However, Yablo argues that, contrary to (*Ex**), even though determinates and their determinables are non-identical properties, some *x*'s possession of some determinable *Q* might be causally relevant to the occurrence of an event *e* even though *x*'s possessing Q's determinate *P* is causally sufficient for *e*. To illustrate, Yablo presents the following two examples. The first involves a pigeon called Sophie who has been conditioned to peck at red. Sophie is presented with a red triangle that she proceeds to peck at. Yablo adds that, as it turns out, the triangle presented to Sophie is not merely red, but scarlet. Since the scarlet colour of the triangle is causally sufficient to produce the pigeon's pecking it follows from (*Ex**) that every other property of the triangle distinct from this property, including its redness, is causally irrelevant. However, Yablo argues that this conflicts with the strong intuition that the redness of the triangle is causally relevant to Sophie's pecking (1992: 257). The second example that Yablo presents concerns properties of events. The buildings in a certain region are causally guaranteed to collapse in the event of a *violent* earthquake, which is any earthquake measuring over five on the Richter scale. A violent earthquake hits the region and the buildings collapse. Yablo adds that it turns outs that the earthquake is not only violent simpliciter, but violent in a specific way – it is *barely violent*, which is to say that it registers between five and six on the Richter scale. Since the bare violence of the earthquake is sufficient to bring about the collapse of the buildings it follows from (*Ex**) that every other property of the earthquake, including its violence, is causally irrelevant. However, once again, Yablo argues that this conflicts with the strong intuition that the violence of the earthquake is causally relevant. Yablo

maintains that if paradigm cases of causal relevance such as these are excluded by (*Ex**) then almost every property is excluded from being causally relevant by (*Ex**). This is because for almost any property that ostensibly looks to be causally relevant to some effect there will be some determinate of that property that is by itself causally sufficient for that effect. For this reason, Yablo says that it follows from (*Ex**) that only "ultimate determinates – properties unamenable to further determination – can hope to retain their causal standing" (1992: 258). However, Yablo argues that even ultimately determinates may turn out to be causally irrelevant according to (*Ex**). This is because even ultimate determinates may incorporate causally extraneous detail that, if abstracted away, leaves a property that would be sufficient for the effect in question.

Yablo argues that all mental-physical relations are a species of the determinate-determinable relation (1992: 256). With respect to mental properties, he says that "mental properties stand to their physical realizations in the relation that regularity bears to squareness, or that colours bear to their shades" (1992: 256). In other words, on Yablo's account of mental-physical property relations, to possess some physical realization of some mental property is to possess that mental property, not simpliciter, but in a specific physical kind of way. Since Yablo argues that the relation between mental and physical properties is a species of the determinate-determinable relation, and that in general some *x*'s possession of some determinable *Q* might be causally relevant to the occurrence of an event *e* even though *x*'s possessing Q's determinate *P* is causally sufficient for *e*, this explains how mental properties can be causally relevant to the occurrence of events despite the causal sufficiency of their physical bases. What I will now argue is that Yablo's reasons for arguing that mental-physical property relations are a species of the determinate-determinable relation present equally good grounds for thinking that the relation between dispositions and their causal bases is also a species of the determinate-determinable relation. In which case, Yablo's argument against the exclusion argument for epiphenomenalism serves just as well as an argument for the causal relevance of dispositions to their manifestations. Yablo's grounds for claiming that mental-physical property relations are a species of determinate-determinable relation are that the latter relation seems to capture precisely what reigning orthodoxy holds about the nature of the former (1992: 254). According to reigning orthodoxy, mental properties are 'supervenient' on the physical:

"(S) Necessarily, for every *x* and every mental property *M* of *x*, *x* has some physical property *P* such that necessarily all *P*s are *M*s" (1992: 254).

The orthodox view also holds that mental properties are 'multiply realizable' by the physical:

"(M) Necessarily, for every mental property *M*, and every physical property *P* which necessitates *M*, possibly something possesses *M* but not *P*" (1992: 255).

However, (S) and (M) are simply more specific versions of clauses (i) and (ii) above. In other words, together (M) and (S) simply entail that all mental properties are 'asymmetrically necessitated' by their physical bases. Since Yablo holds that the determinable-determinable relation is the paradigm of asymmetric-necessitation, he concludes that,

"(D) Necessarily, something has a mental property iff it has also a physical determination of that mental property". (1992: 256)

What reigning orthodoxy says about mind-body relations also applies to the relation between dispositional properties and their causal bases. Dispositions are also said to be supervenient on, but multiply realizable by, their causal bases. For example, it may necessarily be true that anything that has molecular structure *x* also has the property of fragility. However, the converse necessitation does not holds since it is metaphysically possible that the fragility of an object could be realized by some distinct molecular structure, *x**. This point is made by Prior, Pergetter and Jackson themselves in defense of the claim that dispositional properties are distinct from or non-identical with their causal bases (1982: 253). Consequently, if we share Yablo's grounds for asserting (D) then, by parity of reasoning, we should also be committed to,

> *(D*)* Necessarily, something has a dispositional property iff it has a physical determination of that dispositional property.

Yablo thinks not only that mental properties and events can be causally relevant despite the causal sufficiency of their physical bases, but also that mental properties and events can be better candidates for the role of cause than their physical bases. I will end by

suggesting that dispositional properties are better candidates for the role of the cause of their manifestations than their causal bases for exactly parallel reasons. Yablo distinguishes the relations of causal sufficiency and causal relevance from that of causation. He says that something may be causally sufficient for some effect though it incorporates a great deal of causally extraneous detail. Conversely, something may be causally relevant for some effect though it excludes much of what is crucial to the occurrence of that effect (1992: 273). However, Yablo argues that what distinguishes causation from these other two relations is that causes are "commensurate with" (1992: 274), or "proportional to" (1992: 277), their effects. Yablo claims that for any *x* and any *y*, *x* is proportional to *y* if and only if the following four conditions are satisfied. First, y must be 'contingent' on *x*:

> "(C) If x had not occurred, then y would not have occurred either" (1992: 273).

Yablo gives the example of Socrates guzzling the hemlock. He argues that intuitively the better candidate for the role of the cause of his death is the determinable of his guzzling the hemlock, his drinking the hemlock. This is borne out by (C). It is not true that if Socrates had not guzzled the hemlock then he would not have died. However, it is true that if he had not drunk the hemlock then he would not have died.

Second, Yablo says that *x* must be adequate for *y*:

> "(A) If x had not occurred, then if it had, y would have occurred as well" (1992: 274).

Yablo gives the example of a valve mechanism that stiffens due to some freak molecular misalignment and consequently opens too slowly under pressure resulting in the boiler exploding. He adds the additional detail that if the valve had not opened at all then the boiler would not have exploded because the connecting pipe would have burst instead. Yablo argues that in this case we want to say that the cause of the explosion was not the valve opening *simpliciter*, but its opening slowly. This is explained by (A). Though the boiler's exploding was contingent on the valve opening, in a nearby possible world in which the valve did not open slowly it would be true that if it had opened slowly then the boiler would have exploded.

Third, Yablo says that *x* must be required for *y*:

"(R) For all $x^- < x$, if x^- had occurred without x, then y would not have occurred" (1992: 276).

To illustrate the appeal of this requirement, Yablo gives a variation of the example of Socrates's guzzling the hemlock. Supposing that Socrates was a sloppy eater who could not drink without guzzling then it is both the case that Socrates' death was contingent on his guzzling the hemlock and that his guzzling the hemlock was adequate for bringing about his death. Nonetheless, Yablo argues that intuitively the determinable of Socrates's guzzling the hemlock, his drinking it, is still a better candidate for the role of cause. According to Yablo, this is because his guzzling the hemlock was not required for bringing about his death. His merely drinking it would have been sufficient.

Yablo's fourth condition on proportionality is that x must be 'enough' for y:

"(E) For all $x^+ > x$, x^+ was not required for y (1992: 277).

To illustrate the appeal of this final requirement, Yablo gives a variation of the example of the boiler exploding in which the valve stiffens, not due to some freak molecular misalignment, but due to a preexisting structural defect. In a close possible world in which the valve did not open at all it would be true that if it had opened then the boiler would have exploded. Therefore the opening of the valve is contingent, adequate and required for the explosion of the boiler. Nonetheless, Yablo argues that it is still the determinable of the opening of the valve, its opening slowly, which is the better candidate for the role of cause. According to Yablo, this is because the valve's opening simplicter was not enough for the resulting explosion. The valve needed to open slowly.

Yablo argues that with respect to some effect *e*, when faced with a choice between two candidate causes of *e*, *C* and *C**, normally whichever is more proportional to *e* is the candidate to be preferred (1992: 277). According to Yablo, mental events and properties are often more proportional to an effect than their physical determinations. Therefore, in any such case they will be better candidates for the role of cause. He gives the example of acting on one's decision to ring the doorbell, rather than knock. This decision had a physical determination, *p*. However, if in the nearest possible world the physical determination of one's decision to ring rather than knock had been something distinct from *p* then the event of one's ringing the door bell would still have occurred. In

other words, *p* was not required for the occurrence of the act of ringing. However, the act of ringing was contingent on the decision to ring, and the decision to ring was adequate, required and enough for the act of ringing. Therefore, it is the decision that is the better candidate for the role of cause. Similarly, I propose that dispositions are more proportional to their manifestations than their causal bases. The physical determination of some disposition is not required for its manifestation because dispositions are multiply realizable. This makes the physical determination of the disposition, its causal bases, less proportional to the manifestation than the disposition. Substituting 'decision' for 'intention', I believe that Yablo's example of one's intentionally ringing the bell would be a case in point.

5.4 Two Further Objections to the Causal Theory

Having presented my solution to the problem of causal deviance, in this section I address two further objections that might be raised against the causal theory of intentional action, aside from that problem. The first relates to the possibility of spontaneous intentional action. To illustrate this first worry, consider the following example. Suppose that, walking along, someone unexpectedly throws me a ball, which I then spontaneously catch[8]. It might be argued that my catching the ball would be intentional, perhaps on the grounds that catching a ball involves controlled and coordinated bodily movement. However, it might also be argued that I might not have consciously decided or formed an intention to catch the ball. I might have just caught the ball unthinkingly or unreflectively, without any forethought. In which case, this is an example of an intentional action that is not related to any intention to act.

It seems to me that such examples of purportedly spontaneous intentional action do describe actions, but they do not describe intentional actions. Assuming that there really was no moment of decision on my part about whether to catch the ball, my spontaneously catching the ball would be better characterised as sub-intentional action, or as some other form of action that ordinary folk-psychological explanation is unable to adequately explain, and which we need to turn to the concepts of cognitive psychology in order to properly understand and categorize. What reason might there be for saying that it was not intentional? Firstly, the fact that a bodily movement is controlled and

[8] This example is taken from Bratman (1987: 126).

coordinated does not entail that it is intentional. For example, sub-intentional actions may be controlled and coordinated. But one reason for saying that my catching the ball was not intentional relates to Anscombe's thesis, discussed in chapter three, that whenever we do something intentionally we know what we are doing without observation. Anscombe argued that whereas an agent might be surprised to observe that she is doing something unintentionally, such as tapping her foot, or irritating her neighbor with her fidgeting, it is a mark of intentional actions that they are known by the agent without observation. In the example being considered, it is easy to imagine my lacking non-observational awareness of my act of catching the ball. It is easy to imagine that catching the ball might be something that I am surprised to discover I did, in the same way that one might be surprised at one's reflexively reaching out and catching a glass that one spotted rolling off a table in one's peripheral vision, or other instances of so-called 'cat-like reflexes'.

The second worry, aside for the problem of causal deviance, concerns the exercise of bodily skills such as dancing, typing, or playing an instrument. The worry is that many of the things that a person does in the course of exercising such skills and abilities, such as playing C-sharp in the course of playing a sonata, or typing the letter 'i' in the course of typing a sentence, seem to be intentional. However, as with examples of spontaneous action, it will often be the case that the agent never decides or forms an intention to do these things.

One possible response to this objection would be for the causal theorist to simply reject the intuition that actions such as the typist's typing the litter 'i' or the pianist's playing C-sharp are intentional. For example, Setiya characterises these types of actions as "sub-intentional components" of some broader, non-basic action that the agent did decide to do (2007: 55). However, unlike the example of spontaneous action considered above, I find it hard to accept that these sorts of actions are not intentional. A person playing a sonata would never be surprised to discover that she played C-sharp – unless of course this was a mistake. And if the agent were asked retrospectively if she intentionally played C-sharp she would unhesitatingly say yes. She would surely characterise it as something that she did deliberatively.

The problem with the objection, it seem to me, is that it assumes what Bratman calls the Simple View – the view that every action that is performed intentionally is itself specifically intended. However, as already noted, the causal theory does not entail the Simple View. Bratman presents one type of case in which the Simple View appears to be false, which I described above. Actions involving the exercise of bodily skills present another sort of case in which it looks implausible. To illustrate, take the example of playing the sonata. The playing of C-sharp is performed *with* some intention – namely, playing the piece of music of which C-sharp is but one note. In some cases of acting with an intention, the action will itself be specifically intended. Thus in acting the agent will be acting with a *further* intention. In other cases of acting with an intention, such as the playing of C-sharp, it is not plausible to suppose that the action performed was itself intended. Nonetheless, in the latter class of cases it is still consistent with the causal theory to say that the action in question, the playing of C-sharp, was performed intentionally. The reason is that even though the agent did not form an intention to play C-sharp, the playing of C-sharp still manifested *some* intention. This is the intention to execute the broader process of acting - the playing of the piece of music – of which playing C-sharp is a stage, with which the agent was acting. Indeed, this makes good sense on the dispositional analysis of intention. It is true of dispositions in general that, for any disposition *x*, any stage in a process that constitutes a manifestation of *x* itself manifests *x*. This is as true of stages in the process of intentional action as it is of stages in any other processes that manifest dispositions, such as the various stages in the dissolution of a soluble substance, or the various stages in the phototropic movement of a plant towards the sunlight. Intentionally playing C-sharp in the course of playing a sonata, or intentionally typing the letter 'i' in the course of typing a sentence, are merely specific instances of this general feature of dispositions and their manifestations.

There is a further point to note. On a strictly event-causal understanding of causation, causes are events that are temporally prior to their effects. I think that an underlying worry with the causal theory is that it is not clear how this sort of billiard ball causation could apply to the case of the exercise of bodily skills, as if there were one some prior mental event that caused the event of the playing of C-sharp, or the typing of the letter 'i'. However, unlike event causes, dispositions are not causally prior to their manifestations. To illustrate, consider the case of Goalkeeper. Goalkeeper has an intention to guard the goal. If we reject the Simple View then we can say that

Goalkeeper's intentionally catching the ball was not causally related to a specific intention to make that catch, but to the intention to guard the goal with which the catch was made. On an event-causal analysis of the relation between Goalkeeper's catching the ball and Goalkeeper's intention to guard the goal, what caused Goalkeeper to make the catch must have been the temporally prior event of forming the intention. If we accept the dispositional analysis of intention, we need not say this. Rather, the *presence* of Goalkeeper's intention to guard the goal was manifested by each and every catch. It is generally true that dispositions obtain or are present at every moment at which they are manifested.

5.5 Conclusion

In this chapter I have argued that an intention is the disposition of an agent to pursue a goal – a disposition manifested in goal-directed behaviour. This is consistent with Bratman's claim that intentions are associated with propensities not only to act, but also to plan and reason. For planning and reasoning may also be conceived of as manifestations of goal-directed behaviour broadly construed. This proposal also explains how an intention may not only be revealed or exhibited *in* action, but may be future-directed. For dispositions may exist latently without being currently manifested, even without being manifested at all. I have argued that a proper appreciation of the fact that intentions are particular kinds of disposition has important implications. Once we recognize that intentions are a kind of disposition, the possibility of deviant causal chains no longer looks problematic for the causal theory of intentional action. Dispositions can be causes of what they manifest. Further, the manifestation of almost any disposition can conceivably be mimicked. In such cases, there is a sense in which the causal chain is 'deviant', and we explain the sense in which the chain is deviant simply be saying that the disposition is not manifested. Rather, its manifestation is preempted. Once we recognize intentions to be dispositions then the possibility of deviant causal chains ceases to be problematic for the causal theory of intentional action. Typical examples of causal deviance invoked in the literature are simply examples of the mimicking of a special kind of disposition.

Acknowledgements

My thanks to my secondary supervisors, Ian Phillips and Mark Kalderon, for giving me their attention and insightful comments. Especial thanks to my primary supervisor, Rory Madden, for his guidance and support throughout. A version of Chapter Two was presented at a Work In Progress seminar held in the UCL Philosophy Department in March 2011. Chapter four benefited from the helpful comments of the members of a further Work In Progress seminar held at UCL in July 2012. A version of chapter four was also presented at the Berkeley-London Graduate Conference in May 2013 and at the Warwick-London Graduate Mind Forum in June 2013. The suggestions and criticisms that I received on all these occasions were invaluable. This thesis was made possible by the generous support of the Arts and Humanities Research Council and the Royal Institute of Philosophy.

Bibliography

Anscombe, G. E. M. (1957), *Intention*. Oxford: Blackwell

Bargh, J.A.; Gollwitzer, P.M.; Lee-Chai, A.Y.; Barndollar, K.; and Trötschel, R. (2001), 'The automated will: Nonconscious activation and pursuit of behavioural goals'. *Journal of Personality and Social Psychology,* Vol. 81: 1014-1027.

Bayne, Tim (2013), 'Agency as a Marker of Consciousness'. In Andy Clark, Julian Kiverstein and Tillmann Vierkant (eds.), *Decomposing the Will*. Oxford University Press

Bishop, John (1989), *Natural Agency: an essay on the causal theory of action*. Cambridge: Cambridge University Press

Bird, Alexander (2010), 'Causation and the Manifestation of Powers'. In Anna Marmodoro (ed.), *The Metaphysics of Powers: Their Grounding and their Manifestations*. Routledge: Taylor and Francis Group

Brand, Myles (1984), *Intending and Acting: towards a naturalized action theory*. Cambridge, Mass.: MIT Press)

Bratman, Michael (1987), *Intention, Plans, and Practical Reason*. Cambridge Mass.; London: Harvard University Press

Bratman, Michael (1991), 'Cognitivism about Practical Reason'. *Ethics*, Vol. 102(1): 117-128

Bratman, Michael (1992), 'Practical Reasoning and Acceptance in a Context'. *Mind*, Vol. 101 (401): 1-16

Bratman, Michael (1998), 'Toxin, Temptation and the Stability of Intention'. In his *Faces of Intention*. Cambridge: Cambridge University Press

Bratman, Michael (2009), 'Intention, Belief, Practical, Theoretical', in Simon Robertson (ed.), *Spheres of Reason: New Essays in the Philosophy of Normativity*. Oxford University Press

Broome, John (1999), 'Normative Requirements'. *Ratio*, 12(4): 398-419

Broome, John (2005), 'Does Rationality Give us Reasons?' *Philosophical Issues*, Vol. 15(1): 321-337

Brown, Jessica (2008), 'The Knowledge Norm of Assertion', *Philosophical Issues*, 18(1): 89-103.

Charles, David (1989), 'Intention'. In Heil, John (ed.), *Cause, Mind, and Reality: Essays Honoring C.B. Martin*. Dordrecht: Kluwer Academic Publishers

Comrie, Bernard (1976), *Aspect*. Cambridge: Cambridge University Press

Davidson, Donald (1980), *Essays on Actions and Events*. Oxford: Clarendon

Davidson, Donald (1963), 'Actions, reasons, and causes'. *Journal of Philosophy*, Vol. 60(23): 685-700. *Republished in Essays on Actions and Events*

Davidson, Donald (1971), 'Agency'. In Robert Binkley, Richard Bronaugh & Ausonio Marras (eds.), *Agent, Action, and Reason*. University of Toronto Press. Reprinted in *Essays on Actions and Events*

Davidson, Donald (1978), 'Intending'. In Yirmiahu Yovel (ed.), *Philosophy of History and Action*. Dordrecht: D. Reidel. Reprinted in *Essays on Actions and Events*

Davidson, Donald, (1985), 'Reply to Michael Bratman'. In Bruce Vermazen and Merill B. Hintikka (eds), *Essays on Davidson: actions and events*. Oxford: Clarendon

Davies, Martin (1983), 'Function in perception'. *Australasian Journal of Philosophy*, Vol. 61(4): 409-426

Elster, John (1979), *Ulysses and the Sirens: studies in rationality and irrationality*. Cambridge: Cambridge University Press)

Ginet, Carl (2001), 'Deciding to Believe'. In Matthias Steup (ed.), *Knowledge, Truth, and Duty*. Oxford: Oxford University Press

Goldie, Peter (2000), 'Explaining expressions of emotion'. *Mind*, Vol. 109(433): 25-38

Greco, John (2003), 'Knowledge as Credit for True Belief'. In Michael DePaul and Linda Zagzebski (eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: Oxford University Press.

Grice, H.P. (1967), *Logic and Conversation*. William James Lectures: Harvard University

Grice, H.P. (1971), 'Intention and Uncertainty'. *The Proceedings of the British Academy*, Vol. 57: 263-279

Haddock, Adrian (2011), 'That Knowledge That a Man Has of His Intentional Actions'. In Anton Ford, Jennifer Hornsby and Frederick Stoutland (eds.), *Essays on Anscombe's Intention*. Cambridge, MA; London: Harvard University Press

Harman, Gilbert (1976), 'Practical Reasoning'. *The Review of Metaphysics*, Vol. 29(3): 431-463

Harman, Gilbert (1986), *Change in View: Principles of Reasoning* (Cambridge, MA: MIT Press)

Hillel-Ruben, David (1991), 'Naturalized Agency. by John Bishop'. *Mind*, Vol. 100(2): 287-290

Holton, Richard (2008), 'Partial Belief, Partial Intention'. *Mind*, Vol. 117(465): 29-30

Holton, Richard (2009), *Willing, Wanting, Waiting*. Clarendon Press: Oxford

Horn, Laurence (1989), *A Natural History of Negation*. Chicago: University of Chicago Press

Hornsby, Jennifer (2010), 'Trying to Act'. In Timothy O'Connor and Constantine Sandis (eds.), *A Companion to the Philosophy of Action*. Wiley-Blackwell

Hyman, John (forthcoming in 2015), *Action, Knowledge and Will*. Oxford University Press

Hyman, John (forthcoming in 2014), 'Desires, Dispositions and Deviant Causal Chains', *Philosophy* Vol. 89 (1)

Jackson, Frank (1991), *Conditionals*. Oxford: Clarendon Press.

James, Williams (1896), *The Will to Believe and Other Essays in Popular Philosophy*. Norwood, Mass: Plimpton Press

Kenny, Anthony (1975), *Will, Freedom and Power*. Oxford: Blackwell

Kvanvig, Jonathan (2010), 'Norms of Assertion', in *Assertion*, Jessica Brown and Herman Cappelen (Eds.), Oxford: Oxford University Press.

Kolodny, Niko (2008), 'The Myth of Practical Consistency'. *European Journal of Philosophy* Vol. 16 (3): 366-401

Lackey, Jennifer. (2007), 'Norms of Assertion', *Noûs*, 43(3): 594-626.

Langton, Rae (2003), 'Intention as Faith'. In Helen Steward (ed.), *Action and Agency*. Cambridge: Cambridge University Press

Levy, Neil (2011), 'Resisting 'Weakness of the Will''. *Philosophy and Phenomenological Research*, Vol. 82(1): 134-155

Lewis, David (1980), 'Veridical hallucination and prosthetic vision'. *Australasian Journal of Philosophy*, Vol. 58(3): 239-249

Martin, C.B. (1994), 'Dispositions and Conditionals'. *The Philosophical Quarterly*, Vol. 44(174): 1-8

McCann, Hugh (1991), 'Settled Objectives and Rational Constraints'. *American Philosophical Quarterly*, Vol. 28(1): 25-36

McKitrick, Jennifer (2005), 'Are Dispositions Causally Relevant'. *Synthese*, Vol. 144(3): 357-371

Millikan, Ruth Garrett (1996), 'Pushmi-pullyu representations'. In L. May & M. Friedman (eds.), *Mind and Morals*. Cambridge, MA: MIT Press

Moran, Richard (2001), *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press

Moran, Richard (2004), 'Anscombe on Practical Knowledge'. *Royal Institute of philosophy Supplement*, Vol. 55: 43-68

Moran, Richard and Stone, Martin J. (2011), 'Anscombe on Expression of Intention: an Exegesis'. In Anton Ford, Jennifer Hornsby and Frederick Stoutland (eds.), *Essays on Anscombe's Intention*. Cambridge, MA; London: Harvard University Press

Morton, Adam (1975), 'Because he thought he had insulted him'. *The Journal of Philosophy*, Vol. 72(1): 5-15

O'brien, Lilian (2012), 'Deviance and Causalism'. *Pacific Philosophical Quarterly*, Vol. 93(2): 175-196

O'brien, Lucy (2007), *Self-Knowing Agents*. Oxford: Oxford University Press

O'Shaughnessy, Brian (1973), 'Trying (as the mental 'pineal gland')'. *Journal of Philosophy*, 70(13): 365-386.

O'Shaughnessy, Brian (1980), *The Will: a dual aspect theory* (Vol. 2). Cambridge: Cambridge University Press

Papineau, D. (*forthcoming*), 'There Are No Norms of Belief', in *The Aim of Belief*, Timothy Chan (Ed.), Oxford University Press).

Paul, Sarah K. (2009a), 'Intention, Belief, and Wishful Thinking: Setiya on "Practical Knowledge"'. *Ethics*, Vol. 119(3): 546-557.

Paul, Sarah K. (2009b), 'How We Know What We're Doing'. *Philosophers' Imprint*, Vol. 9(11): 1-24

Peacocke, Christopher (1979), *Holistic Explanation: Action, Space, Interpretation* (Oxford: Clarendon)

Pears, David (1975), 'The Appropriate Causation of Intentional Basic Actions'. *Critica: Revista Hispanoamericana Filosofia*, Vol. 7(20): 39-72

Phillips, Ian (in preparation), 'Ignorance of the Future'.

Pickard, Hannah (2004), "Knowledge of Action Without Observation". *Proceedings of the Aristotelian Society*, Vol. 101: 205-230

Prior, Elizabeth W.; Pargetter, Robert; Jackson, Frank (1982), 'Three Theses about Dispositions'. *American Philosophical Quarterly*, Vol. 19(3): 251-257

Raz, Joseph (2005), 'The Myth of Instrumental Rationality'. In *Journal of Ethics and Social Philosophy*, Vol. 1(1): 1-28

Ridge, Michael (1998), 'Humean Intentions'. *American Philosophical Quarterly*, Vol. 35(2): 157-178

Sehon, Scott R. (1997), 'Deviant Causal Chains and the Irreducibility of Teleological Explanation'. *Pacific Philosophical Quarterly*, Vol. 78(2): 195-213

Schaffer, Jonathan (2003), 'Overdetermining Causes'. *Philosophical Studies*, Vol. 114(1/2): 23-45

Setiya, Kieran (2007a), *Reasons Without Rationalism*. Princeton, N.J.: Princeton University Press

Setiya, Kieran (2007b), 'Cognitivism about Instrumental Reason'. *Ethics*, Vol. 117(4): 649-673

Setiya, Kieran (2008), 'Practical Knowledge'. *Ethics*, Vol. 118(3): 388-409

Setiya, Kieran (2009), 'Practical Knowledge Revisited'. *Ethics*, Vol. 120(1): 128-137

Sosa, Ernest (2003), 'The Place of Truth in Epistemology'. In Michael DePaul and Linda Zagzebski (eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: Oxford University Press.

Stalnaker, Robert (1968), 'A Theory of Conditionals'. In *Studies in Logical Theory, American Philosophical Quarterly Monograph Series, 2*. Oxford: Blackwell. Reprinted in Jackson (1991).

Steward, Helen (2012), *A Metaphysics for Freedom*. Oxford: Oxford University Press

Sturgeon, Scott (2008), 'Reason and the Grain of Belief'. *Nous*, Vol. 42(1): 139-165

Thompson, Michael (2012), *Life and Action: Elementary Structures of Practice and Practical Thought*, Harvard University Press.

Velleman, J. David (1989), *Practical Reflection*. Princeton, N.J.: Princeton University Press

Velleman, J. David (1991), 'Review of Michael Bratman's *Intention, Plans and Practical Reason*'. *The Philosophical Review*, Vol. 100(2): 277-284

Velleman, J. David (2000a), *The Possibility of Practical Reason*. Oxford: Oxford University Press

Velleman, J. David (2000b), 'On the Aim of Belief'. In *The Possibility of Practical Reason*

Velleman, J. David (2007), 'What Good is a Will?' In A. Leist (ed.), *Action in Context*. Berlin: Walter de Gruyter, pp. 193–215.

Wallace, Jay (2001), 'Normativity, Commitment, and Instrumental Reason'. *Philosophers' Imprint*, Vol. 1(4): 1-26.

Whiting, Dan (*forthcoming*), 'Nothing but the Truth: On the Norms and Aims of Belief', in *The Aim of Belief*, Timothy Chan (Ed.), Oxford University Press.

Williams, Bernard (1973), 'Deciding to believe'. In Bernard Williams, *Problems of the Self*. Cambridge University Press

Williamson, Timothy (1996), 'Knowing and Asserting', *The Philosophical Review*, 105(4): 489-523.

Williamson, Timothy (2000), *Knowledge and its Limits*. Oxford: Clarendon Press

Wilson, George (1989), *The Intentionality of Human Action*. Stanford, CA: Stanford University Press

Wittgenstein, Ludwig (1953), *Philosophical Investigations*, translated by G.E.M. Anscombe. Oxford: Blackwell

Yablo, Stephen (1992), 'Mental Causation'. *The Philosophical Review*, Vol. 101(2): 245-280

Zalabardo, Jose (2010), 'Why Believe the Truth? Shah and Velleman on the aim of belief', *Philosophical Explorations*, 13(1): 1-21.