Justifying Intentions: Agency, Rationality and Practical Reason

David Olbrich

Submitted for admission to the degree of

Doctor of Philosophy

2020

University College London, Department of Philosophy

Declaration

I, David Daniel Olbrich, 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.

Monday, November 9th, 2020

Impact Statement

This thesis engages directly with two main areas of philosophical work: the philosophy of action, and the theory of rationality. In so doing it hopes to ground further work in a third area: the theory of normativity, particularly in relation to the special kind of normativity that characterizes moral obligation.

More directly, the thesis hopes to aid researchers who wish to shift their focus from the evaluation of actions to the evaluation of plans, intentions, and other practical dispositions.

Abstract

This thesis defends an unusual view within the philosophy of intention: that there are reasons for intention *per se* and that these reasons are not necessarily co-extensive with, or conceptually derivative on, corresponding reasons for action. The question 'what to intend?' is, accordingly, a possible, legitimate and *sui generis* deliberative question, standing alongside the question 'what to do?'. The answering of each of these questions normally involves the answering of the other; though there is this intimate relationship, this should not obscure the possibility or rationality of the free formation of intention for a wider variety of reasons than is usually supposed. Objections to this idea are numerous and important: this thesis addresses, in particular, recent comparisons of intention to belief, particularly the idea that intention aims at good action in the same way that belief aims at truth; reflections on the Toxin Puzzle, sometimes thought to support the inadmissibility to practical deliberation of reasons for intention; conceptions of what sort of attitudes are required for means-end reasoning to make sense; and certain conceptions of the nature of intention that imply that it is answerable only to facts about the worthwhileness of the intended action – such as the conception of intention as itself a normative judgment on action, or as a regulator of action. In contrast, this thesis argues that intention is constitutively a state in which agents take a stand on their own activity, whether that happens through appetitively coming to perform an action in the normal way, or else through the kind of self-control the possibility of free formation of intention offers.

Acknowledgements

It is sometimes easy to forget, while lost in a labyrinth of one's own making, that philosophy is a conversational discipline. Research tasks of this size and scope are difficult precisely because one can become engaged not in communicating a point of view so much as constructing a bureaucracy of ideas, where prior assumptions are to be minimized, results to be rich, extant controversies to be stayed neutral on (the philosophical equivalent of "redirecting to another department"), while everything else becomes gradually subordinated to the all-consuming needs of The Account. We are to be thankful, then, that as philosophers we are never without that red thread that leads us back to safety and sanity. We only have to turn to each other.

My main thanks are due to the three principal supervisors I have had over the many years I have been at UCL: Doug Lavin, Mike Martin, and Ulrike Heuer, chronologically. Doug I found to be uniquely brilliant, and his winning combination of enthusiasm for ideas and undauntedness while thinking about them made him in my eyes a living exemplar of the promise of philosophy. He was a tremendous influence on me, and I hope that I will eventually prove worthy of the investment he made in me when he gave me, as supervisions, huge amounts of his time that went far beyond any plausible obligation he might have had to me – not to mention the restoring effect that every one of my conversations with him had.

My supervisions with Mike had a more concentrated effect on the quality of my work than any intellectual experience before or since. Mike perfectly navigated the balance between encouraging my originality while weaning me off the bad ideas that it frequently produced. I have never learned so much just by listening and trying to keep up, but also never benefited from the same intensity of interest in and insight into the overall shape of my views, including not only my ideas for what to write my thesis on, but also how they related to my long-term preoccupations.

It was Ulrike who enabled me to actually write this thesis. Every chapter that follows was written under her aegis. Her feedback throughout this process – for me the most difficult stage of research – has been invaluable, and she did

more than anyone else to bring me out of my philosophical solipsism and back into the project of engaging with other thinkers in a sustained and disciplined way. Despite my failures on this score, she never dismissed my ideas or interpretations, and always pushed me to do better. I hope that this thesis lives up to the standards she set.

I also owe thanks to Mark Kalderon, who briefly supervised me and helped with a (now defunct) version of a thesis plan, and who also reviewed some of my other research on W.D.Ross. Lucy O'Brien has also been consistently helpful and encouraging.

The graduate community of UCL's Philosophy Department deserves its own recognition. Its collegiality and hospitality helped me, more than I am probably capable of articulating, and hopefully rubbed off on me. I must own up to some obnoxious tendencies, but I must also thank my colleagues for helping to curb them, without ever losing their humour, respect, gentleness, or their time for me. Of the older generation of graduate students who very much set the standard for me to aspire to, I would like to thank Alex Geddes, Henry Clarke, Alec Hinshelwood, Ed Nettel, Ed Lamb, and Lea Salje – the regular common room denizens who helped me feel less lost. Of those who were more roughly contemporary with me, there are too many to name and to thank, but I would like in particular to name Julian Bacharach, Edgar Phillips, Charles Jansen, Laura Silva, Vanessa Carr, Ashley Shaw, Catherine Dale, James Laing, Rowan Mellor, Pete Faulconbridge, and Michael Markunas. I would also like to thank the UCL Philosophy administrative staff, in particular Richard Edwards, for his helpfulness and efficiency in supporting all of us.

I must also thank all my friends (too many to name; but they know who they are) for their support and their receptiveness: in particular I would like to single out Cat Brooks, Sean Whitton, Sumil Thakrar and Max Goplerud. I would also like to thank my mother Danielle and her partner Julian for extensive material support. Finally, this thesis would not have been possible without my analyst Dr Lima Barreto, who brought me back into the world.

Contents

Introduction

In thinking about practical thought, it is natural to assume a hierarchical picture of the relationship between particular actions and the ends which they serve. At the top of this hierarchy are whatever ends we are prepared to consider ultimate; below them, and justified by them, are the actions and ends that contribute to their furtherance; below them stand the same; and so on, all the way down to the ordinary actions of life, which, thanks to this analysis, are now shown to be justified in the light of fundamental values or ends. This picture is everywhere in practical philosophy. Its classic expression is in Aristotle:

> Where there are ends over and above our activities, in these cases the products are by their nature better than the activities. Since there are many sorts of action, and of expertise and knowledge, their ends turn out to be many too: thus health is the end of medicine, a ship of shipbuilding, victory of generalship, wealth of household management. But in every case where such activities fall under some single capacity, just as bridle-making falls under horsemanship…in all activities the ends of the controlling ones are more desirable than the ends under them, because it is for the sake of the former that the latter too are pursued.[1]

This thesis emerges from an increasing suspicion of mine that this picture is strongly incomplete in relation to important sorts of action, and it aims to provide a groundwork for an alternative (or at least supplementary) way of thinking about rational agency. The issue is that not every action can be adequately fitted into a compelling enough hierarchy of this kind. This is not because of the difficulty of interpreting the upper elements of the ends-scale (a difficulty that is central to the Aristotelian paradigm), but rather because certain actions are performed with a view to ends that they simply do not either exemplify or further.

For a relatively trivial example, consider the following instance of playing a game[2]. Suppose an agent plays football in the evenings in order to relax. In playing the game, they will aim to score at least some goals. It need not be true that the more goals they score, the more they will relax, and it certainly

[1] Aristotle 2002, I.1, 1094a5-17
[2] Cf. Nguyen 2020, ch. 1, p.8-9

need not be true that the project of relaxing requires them to successfully score goals; it is enough for that that they play the game. Therefore, in aiming to score goals, they have at least one aim that itself is not instrumentally required by their broader purposes – because scoring a goal neither exemplifies relaxing nor furthers it. What is clearly going on, rather, is that relaxing requires them to enter into the spirit of the game; having entered into the spirit of the game, the agent aims at scoring goals; they now pursue that accordingly. This perfectly natural and intelligible progression goes beyond the hierarchical picture of justification of actions that is centred on means and ends, yet the end result is that the agent's specially acquired aim is vindicated though its relation to these ends.

Such cases bear some resemblance to those at play in traditional discussions of two-level justification in the literature on consequentialism[3]. There too the justificatory relation is indirect and to some extent psychologized: there is one set of justifying considerations that recommends the adoption of certain dispositions (such as the disposition to love and to favour), a disposition which itself will often manifest in ways unjustifiable by the light of these very same considerations (because they lead to favouring some people over others in ways contrary to the impartial spirit of those considerations). But we don't need to be consequentialists in order to find something that looks like a two-level justification structure, and such cases do not necessarily invite consequentialist theorizing.

Putative examples of two-level structures abound, particularly in the context of professional work. Williams suggests the following in the case of lawyers[4]: that some lawyers who enter into the spirit of defending their client may (depending on the client) end up acting in ways that are fairly horrible, and contrary ultimately to the interests of justice if they are successful, yet it would be inappropriate for them to tone down their behaviour, against their client's legal interests, in order to produce the outcome they personally judge would best serve justice. This is because the adversarial system leaves such

[3] The classic presentation of such a view is in Hare 1981. An application of such a view to the ethics of partiality towards others is contained in Ashford 2000, section IV.
[4] Cf. Williams, 'Politics and Moral Character', p.64-5, in 1981; 'Professional Morality and its Dispositions', in 1995.

judgments up to the court, and it cannot function as it is designed to unless each advocate presents the strongest possible case for their client so that the court may adjudicate. If the effective delivery of justice really does require this sort of adversarial system, then at least some participants in that system will have aims (i.e. to win certain cases for the wrong side) that cannot be justified as part of furthering justice, yet justice also helps to vindicate their having these aims. Here, again, the hierarchical picture cannot quite apply, because such actions lack the normal sort of means-end justification that that picture is built on.

It should be obvious that such cases do not only introduce an apparently unusual scheme of justification, but also centrally depend on *the agent's participation* in the relevant practice. It is by entering into the spirit of some practice that the agent gives themselves aims that are non-instrumentally vindicated either by the considerations justifying their participation in that practice, or by the considerations justifying the existence of the practice itself. These aims are not derived from the broader justifying considerations through some principle of practical reason, but rather enter onto the scene through the agent's active participation in a practice (sometimes in a pre-existing social practice, but not necessarily, as in the cases of friendship and love). In participating in this practice, the agent gives themselves these new aims, because the adoption of such aims is what the practice requires, or at least invites. This is not just a genealogical point: it is the agent's continuing active participation that keeps these aims going. When the evening footballer stops playing, they stop aiming to score goals; if the lawyer abandons the case, they will stop valuing the legal interests of their client; if a friend ceases to value a friendship, they will stop aiming at interacting with their former friend in making decisions about what to do with their time and energy.

So in understanding such cases, we have to focus on *higher-order agency*: exercises of agency which consist in an agent's putting themselves into, and keeping themselves in, states that are themselves characterizable through what sort of agency they involve – just as the footballer actively participates in the game and, within that participation, possesses further aims non-instrumentally justified by whatever justifies participation itself. If higher-

order agency is possible, then it is possible to put oneself into a psychological state from which further actions of one's own will spring. By exercising higher-order agency, we put ourselves into a state that determines which actions we will subsequently perform.

Within extant philosophical literature, the concept of higher-order agency has been explored most rigorously in the context of the theory of decisions and intentions. While the detailed structure of the issues differs from that described above, the element of higher-order agency is, I will now argue, shared. This is because decisions as to how to act in the future put us into states (intentions) which in turn lead to actions. Pink (from whom I draw the term 'higher-order agency'[5]) writes:

> we have two forms of action control. Not only do we have and exercise control over our actions as we perform them. We also have a control over our actions which we can exercise in advance. By taking a decision now to go to the dentist this afternoon, I can ensure that from that time onwards I intend or remain decided on going - an intention which persists to the time of action, so that I eventually go. On the other hand, had I decided not to go to the dentist this afternoon, that decision would have prevented me going. I should have formed a persisting intention not to go to the dentist. So, given its effects on subsequent action, decision-making can be used to exercise control in advance over which future actions the decision maker performs. Decision and intention control gives us future action control.[6]

Ordinarily, of course, the relation between decision and action does not normally exemplify the discrepancy in justification noted above for participation in practices. Normally, when we decide to act in some particular way, it is because we are motivated to act in that way: we don't normally think of our decisions as justified differently to how we would justify doing what they are decisions to do. In contrast, we often think about our participation in practices as justified in ways that don't pertain to the aims we have within that participation – and this suggests that the higher-order agency of practice participation might function differently to the higher-order agency of decisions. But if that's so, we need an account of what exactly the difference is.

---

[5] E.g. Pink 1996, ch. 1 p.22
[6] Ibid.,p.16

Our focus, then, is not just on higher-order agency, but specifically on the sort of higher-order agency that seems to be involved in participation in practices: that is, free and deliberate higher-order agency. In the context of the theory of intention, this is the possibility that decisions may be justifiably and intelligibly taken for reasons unrelated to whatever they are decisions to do, without offence to the rationality of the agent taking these decisions. Further argument (which I will not attempt here) might establish that the higher-order agency of practice participation is an *instance* of the higher-order agency of decisions: that adopting the aims that deliberate participation in a practice requires us to adopt is possible on account of the way agents are able to aim at φ-ing on the basis of having deliberately taken a decision to φ for reasons unrelated to φ-ing itself[7]. The hope (unrealized within the scope of this thesis) is that a very general account can be given of the relationship between the rationality of actions and the rationality of the practical dispositions to perform those actions, even where those dispositions cannot be characterized in terms other than what they are dispositions to try to do.

However, the idea that higher-order agency can be freely and rationally exercised in this way is not the usual position – quite the opposite. As we will see, it is often thought that the rationality and intelligibility of intentions is subordinate entirely to that of the actions which they are intentions to perform, so it is only in virtue of some action's being a sensible action to perform that some intention is a sensible one to have. I will explore the opposite position, that intentions can be rationally chosen for reasons that just reflect the usefulness of the intention and not of the actions they are intentions to perform (perhaps in some cases the agent does not even need reasons in order to form an intention); and that an agent's adopting an intention on such a basis reflects a perfectly normal aspect of agency. This is roughly the position defended here (some qualifications to the rationality of free higher-order agency are conceded).

[7] This would involve interpreting the decision to participate in a practice as, *inter alia*, a decision to (try to) do what that practice requires its participants to try to do, and so invites discussion of what practices constitutively are – a topic tangentially related to the theory of intention.

The idea that an agent's decisions are just as much up to them as their actions has often been objected to. A flat-footed objection is made by Brian O'Shaughnessy, who states:

> there is no activity of 'deciding to do φ', where φ is an act; even though there is an event of deciding-to-φ. This receives confirmation in the fact that there is no order: 'Decide to raise your arm.'[8]

But the basic observation here is dubious. It is surely possible to decide now to raise one's arm in three minutes' time, and even to make that decision intentionally and in response to an order. What would make a little less sense is the order: 'decide to raise your arm now', but that just reflects the fact that the way to make that particular decision is to try to raise one's arm. The decision itself is not obviously distinct from the activity, at least in terms of what the agent can do. Thus the unintelligibility of the order 'decide to raise your arm now' seems to reflect only on present-directed decisions and only in virtue of referring in a confusing way to know-how that has nothing to do with the distinctive content of the order. It is not an argument against the agent's active power to make decisions that in certain cases to make such a decision is not distinct from trying to do what it is a decision to do. Furthermore, as I will urge, the ability to make decisions actively (even decisions to do particular things, as opposed to the ability to make up one's mind one way or the other) itself need not be thought of as arcane, let alone a philosophers' nonsense: at least some situations call for the ability to make such decisions because such decisions are good decisions to make. If that's right, then the real task is to correct philosophical theory in light of this ordinary fact.

The ability to make decisions on the understanding that one will then act as decided (unless one changes one's mind, or is interrupted, and so on) constitutes an ability to act in a way that ensures that further actions will happen from that action. This is exactly the phenomenon I have been describing so far – that is, the higher-order agency of decisions: agency that consists in psychological changes that themselves will generate further actions.

---

[8] O'Shaughnessy 1980b, p.300

Part 1 of the thesis is devoted to a basic argument for, and defence of, higher-order agency in the context of intention – of the possibility of intentionally intending to do something, and in particular, the possibility of its *rationality*. The arguments of two philosophers, Nishi Shah and Pamela Hieronymi, suggest that intention is something more like a way of correctly processing information: it is something we get to if and only if we properly think through our reasons for action, not just as a normal psychological consequence of such thought, but rather because it is the constitutive aim or standard of intention to be governed by reasons for action. These arguments are rebutted in chapter 1.

In chapter 2, I introduce the idea of higher-order agency with reference to clear illustrative cases that show that the capacity to intentionally adopt and maintain certain intentions constitutes a useful extension of our practical powers as agents: that there are cases in which there is reason to adopt certain intentions irrespective of whether there are corresponding reasons to perform the actions that would be intended. These cases are built to be counterexamples not only to the very restrictive claims of Shah and Hieronymi, but also to a more expansive claim made by Pink, on which there are reasons to decide to φ that don't correspond to reasons to φ, but only if, after that decision is made, there is just as much reason to φ as there is to intend to φ.

A focal point of objections in the philosophical literature to the rationality of higher-order agency is Kavka's Toxin Puzzle. The examination of such objections is carried out in chapter 3. I argue that the Toxin Puzzle does not show that there are no reasons for intention *per se*. At most it shows that there is a rational requirement on agents not to adopt intentions when they are certain that they won't do what they intend. This requirement is shown to be wholly consistent with the possibility of rational higher-order agency; at most it prohibits the formation of rational intention in a fraction of cases.

The next section of the thesis defends the possibility of higher-order agency with attention to more abstract theoretical issues. It tackles, on its own terms, a fundamental conception of intention as a state whose constitutive function

is to apply our judgments about what we should do to our actions, resulting in our acting in conformity with those judgments. That function is central, but not constitutive. Chapter 4 makes a positive argument against the idea that intentions involve 'settling' what to do. It shows that this claim renders unintelligible the old idea that action is the conclusion of deliberation. Furthermore, it is vulnerable to counterexample. Though having intention involves something in the way of a positive answer to the question 'what to do?', it doesn't close that question. At most it sets up a *presumption* of what the answer shall be: a glimpse of the will in action.

Chapter 5 prepares the groundwork for a positive theory of intention that avoids the implication that intention is governed only by reasons pertaining to the question 'What to do?'. It argues that the notion of teleological agency, of the capacity to act for the sake of an end, must be treated as conceptually prior to the notion of intention, so that intention is understood in its terms. In this respect, intending must be treated as similar to trying: the priority of agency is requisite if we are to avoid a certain background inconsistency without succumbing to the pitfalls that await views that deny any of the attractive propositions that together make up the inconsistency.

These two strands are drawn together in chapter 6, which describes the *stand-taking* conception of intention. To intend to act is to exercise one's agency by taking a stand on what one shall do. Yet the idea of intention as an exercise of agency (but not an action) needs to be filled out. This chapter argues that it helps with explaining some of the fundamental rational constraints on intention: the predictive requirement not to intend what can't or definitely won't happen, and the normative requirement not to intend certain actions that are repugnant or have nothing positive going for them. The stand-taking conception, unusually for conceptions of intention, is consistent with the rationality of adopting intentions for their own sakes, which is why its elaboration in this chapter is crucial to the overall project.

Chapter 7 defends the stand-taking conception from an important objection: that it is inconsistent with the following idea within the theory of instrumental rationality, that taking the means to one's ends is justified only if the ends

themselves are justified, and consequently, valid instrumental reasoning can take place with respect to some end only if the agent takes the achievement of that end to be justified. If this is right then it matters that ends are justified and not just that intentions to achieve them are justified. It is argued that the idea explained in chapter 6, of intention as an exercise of agency, solves this problem.

1. Initial arguments against the possibility of state-given reasons for intention

As explained in the Introduction, the aim of this thesis is to make room for an understanding of rational higher-order agency: agency that consists in putting oneself into states in which agency of a further sort is exercised. This formulation is itself unclear for the time being – and its clarification is the task of much of this thesis, especially chapters 5 and 6 – but it simplifies things to understand the core aim as being to understand how an agent could quite sensibly adopt an intention to perform some action on account of something other than the perceived merits of that action itself.

In turn, this simplification disguises some important aspects that are buried in the more obscure formulation. In particular it glosses over the relationship between intention and action itself: between the state that is the object of higher-order agency, and the events that are the upshot of that state. Intentions to act are the kinds of things that lead to actions: without that, we lose our grip on what we're talking about. It would be particularly strange if it were possible for an agent to rationally adopt some intention and yet not be able rationally to execute that intention. For example, suppose someone makes a decision to try out a new pub that has just appeared on their street. The terms in which we might naturally criticize that decision – 'but the risk of spreading the virus is too great!' – are exactly the terms in which we would criticize that *action* of going to that pub. If the action is criticisable, it seems, so is the decision. So if that decision *is* justified, then it cannot be wrong to try out that pub. Were it wrong, then the decision wouldn't have been justified – or so it seems. (In this thesis, I understand decisions as identical to acquisitions of intention, or initiations of intention, staying neutral on whether we should understand these as themselves actions.)

It is therefore plausible that there is a tight connection of some kind between what justifies an intention to act and what would justify that action itself – and also, between what would justify not making that decision and what would justify doing something else. In the following chapters (1 through 3) I

argue that this connection has been conceived too tightly in contemporary philosophy. I shall call the kind of position I criticize the 'good-action norm' position: it is the position that what justifies an intention to act is exactly, and only, what would justify that action itself.

Many conceptions on which the justification of intentions is characterized by the good-action norm do not explicitly argue for the correctness of that norm. Those that do will be the subject of most of the discussion in what follows; the purpose of this opening section is to outline the main ways in which this position is, as it were, adopted by default – and to deflate any sense of disagreement with those conceptions by showing how they can be easily modified to avoid entailing the good-action norm by default, or else by showing that their core concerns are not in tension with the existence of intentions that are justified on merits other than those of the actions they are intentions to do[9].

First, a little more terminology. Following Parfit and later philosophers[10], I shall use the term 'state-given reasons' to refer to all those reasons to have mental states that do not bear on the object of those states by showing it to be true, or justified, or whichever term of evaluation is relevant. For example, a financial reward for having a certain belief is a state-given reason to have it, because it is not a reason that justifies a belief by showing how its object is true – which is the relevant mode of object-based evaluation for beliefs. A financial reward for adopting an intention is a state-given reason for that intention because it does not bear on the worthwhileness of doing what it is an intention to do – what might be thought of as the object of that intention. This way of setting up the distinction is essentially contrastive, and more, contrastive against philosophically articulated restrictions on reasons: 'state-given reasons' is a way to refer to what have been thought to be deviant

[9] I shall pass over the work done in decision theory on these issues, though some work has been done very recently (cf. in particular Hedden 201; Bales 2020, in defence of conceiving of the agent's decision-theoretic options as their possible decisions rather than their possible actions) as well as previously (cf. in particular Gauthier 1986; Kavka 1978). As far as I know, there is within decision theory no explicit argument *for* the good-action norm or even any particularly deep conception of what it is, though plenty of arguments against it, so it is not within the scope of this chapter to evaluate.

[10] Parfit 2001 p.22-3; Schroeder 2012, esp. Section 1B.

reasons, reasons that do not really work to justify – in just the way that certain beliefs, no matter how well-rewarded they are, are not justified if they are known not to be true. (They may of course be good reasons for related but distinct states, such as the state of desiring to have the belief for which there is state-given reason). In using this terminology, I do not mean to suggest that state-given reasons for intention really are deviant; I use it rather to mark the distinctive sorts of reasons whose genuineness I aim to establish.

In the modern Humean tradition, some theories of the nature of intention as a psychological state lead to the good-action norm. In particular the exclusion of state-given reasons is an implication of Davidson's theory of intention, on which intention is conceived as a judgment that the action it is an intention to do ought to be performed[11] (such views regarding the nature of intention are explored more in chapter 5). If intention is just such a judgment, then, just as with beliefs in general, it is rationally controlled only by evidence pertaining to the truth of that judgment: that is, evidence in favour of the proposition that the action ought to be performed[12]. Once an agent has such evidence, then they can rationalize to themselves the performance of that action, and so intelligibly act intentionally. State-given reasons will therefore be disallowed; only reasons for action, that act as evidence for this sort of 'ought' proposition, are admissible to practical deliberation on this view.

In Davidson's case, the identification of intention with judgment occurs as part of a series of similar identifications: of desire with pro attitudes, of pro attitudes with judgments of evaluative sentences as true[13], of these with

[11] 'In the case of pure intending, I now suggest that the intention simply is an all-out [evaluative] judgement. Forming an intention, deciding, choosing, and deliberating are various modes of arriving at the judgement, but it is possible to come to have such a judgement or attitude without any of these modes applying.' "Intending", in *Essays on Actions and Events* (2001), p.99.

[12] A suggestion explored in more detail in "How is Weakness of the Will Possible?", in ibid.

[13] 'someone who says honestly 'It is desirable that I stop smoking' has some pro attitude towards his stopping smoking. He feels some inclination to do it; in fact he will do it if nothing stands in the way, he knows how, and he has no contrary values or desires. Given this assumption, it is reasonable to generalize: if explicit value judgements represent pro attitudes, all pro attitudes may be expressed by value judgements that are at least implicit.' Intending, in ibid., p.86

desires and wishes[14] and more mysteriously even actions themselves[15]. But the identification here of desires, wishes, pro attitudes and evaluative judgments is not necessary for a way of thinking about action in terms that make the good-action norm seem natural. It also gains some traction from the basic account of action as rationalized in the light of the agent's beliefs and desires together with a *non*-judgmental conception of desire. Desires too are sometimes thought to be rationalized only in virtue of features of their object and not by anything like a state-given reason[16]. If this is correct, and if intentions are desires, then intentions too will not admit rationalization by state-given reasons: only object-related reasons will matter.

The exclusion of state-given reasons from the justification of intentions is therefore natural, though perhaps not inevitable, if intentions are identified with desires, but also if they are identified with beliefs. It is not inevitable since it is open to a theorist to assert that just those beliefs or desires with which *intention* is identified are justified in the light of special norms, or without the usual constraints. Velleman, though not a Humean, is an example of a philosopher who prescribes special treatment of this sort for intentions (which he suggests are kinds of predictive beliefs, a view discussed in chapter 5). He suggests that the kinds of beliefs which are intentions are able to be formed, unlike normal beliefs, voluntarily:

> the agent has an unusual amount of discretion in making these predictions. Until he forms an expectation about what he's going to do, he isn't going to do anything…The agent's reflective reasoning thus leads him to optional, self-fulfilling predictions of acting that are regarded as such. I can see no difference between such predictions and intentions.[17]

---

[14] 'We may put aside wishes for things that are not consistent with what one believes, for these are ruled out by our conception of an intention. And we may put aside wishes that do not correspond to all-out judgements…But once we put these cases aside, there is no need to distinguish intentions from wishes. For a judgement that something I think I can do, and that I think I see my way clear to doing, a judgement that such an action is desirable not only for one or another reason but in the light of all my reasons; a judgement like this is not a mere wish. It *is* an intention.' Ibid., p.101

[15] 'the fact that the action is performed represents a further judgement that the desirable characteristic was enough to act on—that other considerations did not outweigh it. The judgement that corresponds to, or perhaps is identical with, the action cannot, therefore, be a prima facie judgement; it must be an all-out or unconditional judgement which, if we were to express it in words, would have a form like 'This action is desirable'.' Ibid., p.98

[16] E.g. Scanlon 1998, p. 38

[17] Velleman 1985, p.53-4

According to Velleman, though these predictions have factual content, they are unique in not being governed solely by evidence as to their truth; rather agents are supposed to have some freedom with respect to these predictions that lies in the fact that if the agent makes the prediction then it will be true. If there is an inconsistency here with the possibility of state-given reasons for intentions, or such predictions, then it is one of a more subtle kind; it is not built into the shape of the view.

The possibility of such a theory which allows for intention-constituting beliefs and desires to be pragmatically adopted, perhaps adopted even in the light of state-given reasons, suggests that the good-action norm is not a strict consequence of a theory of intention as either a kind of belief, desire, or some combination of the two. Though any such position is vulnerable to the criticism of being *ad hoc*, it is certainly not automatically incoherent.

Apart from these psychological views, the partial exclusion of state-given reasons from the justification of intention is also found in literature focused on the guise of the good: the idea that all actions are performed either with, or more usually from, an apprehension of their merit[18]. Every action that is performed from an intention will therefore also be performed from a perception of its perceived merit. In the first instance this is only a theory about (intentional) action itself, not a theory of intentions, so the way in which the former constraints the latter depends on additional assumptions that connect intention to action. One might try to separate the two topics, so that whatever we wanted to say about the possibility of state-given reasons for intention would not affect our theory of action.

However, intention and action are plausibly connected – and that means that the guise of the good is a significant thesis for the theory of intention as well as the theory of action. The relevant assumption is that if it is rational for an agent to intend to act some way, then it is rational for them also to act that way[19]. This view is an implication of (but does not in turn require) a conception of intention as the mediator between judgments of how to act, and

[18] E.g. Tenenbaum 2020; Raz, "Guise of the Good", in Tenenbaum (ed.) 2010.
[19] The topic is extensively explored in Bratman 1987, chapter 6.

actions themselves, responsible for ensuring that the agent performs an act that they judge meritorious. Pink describes this view as ascribing a 'reason-applying' role to intention (here reframed in terms of decisions, which he conceives as actions that initiate intentions):

> the will is an *executive* capacity - a capacity for *applying* our prior deliberations about how to act. A decision is a second-order executive action - an action by which we ensure that we subsequently perform the first-order actions which, as deliberators, we have judged it desirable to perform…We must be able to exercise control in advance over what future actions we perform - a control that we exercise through a prior action-generating agency of decision-making.
>
> Our ordinary conception of the will is, I argue, precisely a conception of it as a faculty for second-order executive agency – an agency the function of which is to apply reason as it governs the first-order agency of the actions decided upon. For, as we ordinarily conceive it, decision-making provides a highly distinctive, *reason-applying* way of controlling one's future actions.[20]

Any theorist who thinks that actions are properly controlled by an agent's judgments about which action is desirable, and that intention is part of the normal psychological route by which such judgments translate into actions, is likely to agree with this conception. Since agents act from their intentions, an agent who possesses an intention to do something undesirable will go on to do something undesirable (assuming that it is something that they are capable of doing, and nothing stops them, and so on); an agent who possesses an intention to do something that it would be irrational for them to do will (normally) go on to act irrationally. It does not take much to push intuition in accordance with this. Rational intentions to do irrational things involve their agents in failures of rationality. If there could be a rationally *required* intention to do an irrational thing, then an agent could end up in a situation where they are irrational whatever they do: if they don't form that intention, they are irrational, but if they do, they will act irrationally. If one is attracted to the idea that full rationality must be achievable by an ideal agent in non-ideal circumstances, then this may well be thought to preclude such possibilities, since such situations are ones in which no agent in principle could succeed in being fully rational.

[20] Pink 1996, Introduction p.5

If this assumption is correct, then it bears on the intelligibility of state-given reasons for intention, precisely because state-given reasons seem to make such rationality-defying situations possible. Suppose that state-given reasons really do make it rational to adopt certain intentions; then it could conceivably be that the applicable state-given reasons for some intention are weighty enough to ground its being rationally required even if the object of that intention is itself irrational. In that case, an agent would be obliged to adopt that intention for those reasons, but then they will act irrationally. If such situations are deemed impossible in principle, then state-given reasons for intention may have to be in part rejected: though perhaps some intentions to do rationally permitted things may be justified through state-given reason, it will have to be made impossible for intentions to be justified by state-given reasons even where the intended action is contrary to the agent's reasons. A form of this worry motivates Pink to posit limitations on what kinds of state-given reasons for intention exist[21].

Though the discussion of this issue has so far been very abstract, in chapter 2 I will argue more concretely that there *are* conceivable situations in which an agent is rationally required to adopt some intention even though the associated action is worse than some of the available alternatives. Assuming as well that rational intention does guarantee rational action, this means that the guise of the good itself must be heavily qualified: at a minimum it cannot imply that, for rational action to take place, the agent must take the intended action to be the best of the available options.

There is an even more pressing issue for the possibility of state-given reasons for intention in light of the guise of the good: the possibility of justified intentions to perform actions in which the agent can see *no* merit. It is not to the point here that such action are worse than some of the agent's available alternatives, but that if they were to perform such an action, they could not do so from a perception of the action's merit, simply because the action has nothing going for it. The usual reference point for the idea of acting without seeing any merit in one's action is Quinn's example of the 'radio man':

[21] Cf. Pink, ibid., ch. 6, section 'Reason-Apply'.

> Suppose I am in a strange functional state that disposes me to turn on radios that I see to be turned off. Given the perception that a radio in my vicinity is off, I try, all other things being equal, to get it turned on…in the case I am imagining, this is all there is to the state. I do not turn the radios on in order to hear music or get news. It is not that I have an inordinate appetite for entertainment or information. Indeed, I do not turn them on in order to *hear* anything. [22]

In Quinn's example the radio man is possessed by a brute 'functional state', not anything like an intention. It would manifest as a series of attempts to get radios turned on, without any accompanying thoughts as to the point of that. Quinn draws the general moral that rationalization of what is done requires evaluative concepts:

> A noncognitive pro-attitude, conceived as a psychological state whose salient function is to dispose an agent to act, is just not the kind of thing that can rationalize. That I am psychologically set up to head in a certain way, cannot by itself rationalize my Will's going along with the setup. For that I need the thought that the direction in which I am psychologically pointed leads to something good (either in act or result) or takes me away from something bad.[23]

If this is correct, then an intention rationally adopted for state-given reasons cannot rationalize the action intended if the agent does not see any point in that action. On the assumption that rationally adopted intention ensures rational action from that intention, it follows that the intention could not have been rationally adopted in the first place: that state-given reasons cannot justify unless appropriate object-related reasons are also in the picture.

The correct response here for the theorist who wishes to defend the unrestricted applicability of state-given reasons for intention is to dispute that an intention rationally adopted for state-given reasons wouldn't thereby rationalize an action done from it. In Quinn's case, the radio man's behaviour is obviously unintelligible to themselves. But suppose we take a different, artificial case, where someone has formed an intention to do what the radio man does, because they have been offered a reward if they possess that intention for some amount of time. As a consequence, they act from that intention, but without seeing any merit in their actions, beyond the fact that performing that action is a *sine qua non* of maintaining the continued

[22] Cf. Quinn 1980, p.236-7.
[23] Ibid. p.242

possession of a desirable intention. However, just focusing on the production of the action by their intentions, the agent of this modified case is not in as radical a position of Quinn's radio man, who is capable of interpreting their behaviour only through the lens of 'brute functional disposition', and not only because they have to perform the action in order to continue to possess the rewarded intention.

This agent has another way of interpreting their actions: as the product of a justified intention. This is not a *brute* functional disposition: their actions are the product of what we are supposing is a rationally adopted intention. So even if the agent sees no merit in their action (besides its being a *sine qua non* of the possession of the justified intention) there are no grounds here for denying its intelligibility or its status as an action. *Qua* product of a rational intention, the action is understandable. This is not to say that *all* actions that would be the upshots of intentions for which there is significant state-given reason would be justified; discussion of the Toxin Puzzle, in particular, occurs in chapter 3. But the guise of the good cannot be proved in sufficiently strong terms so as to rule out the possibility of the intelligibility of this sort of state-justified intention. Examples such as Quinn's do not engage with what an agent's perspective on their actions would be were they to understand them as the product of sensible intentions; they cannot disprove the intelligibility from the inside of such actions.

The guise of the good thesis does present a more significant barrier to the possibility of unrestricted rational higher-order agency than does the identification of intentions with desire or judgments. It is a more important objection to the idea that intentions can be rationally adopted for state-given reasons. More weighty arguments are addressed in what follows.

Intention and Belief 1: Shah on practical and doxastic deliberation

The overall theme of the arguments examined next is that intention functions similarly to belief, whether in terms of its constitution, the forms of normativity appropriate to it, or even in the kinds of philosophical arguments that can be brought to bear on such questions. The one we shall examine next

sets up a desideratum for any conception of intention that we essentially accept, even though we reject its conclusions. This is the comparison made implicitly by Nishi Shah in a series of papers but centrally in two papers with very suggestively similar titles, 'How Truth Governs Belief' (2003) and 'How Action Governs Intention' (2008). These papers do not *assert* that belief and intention function similarly; rather, in attempting to structure discussions of intention and belief in similar ways, dealing with positions and objections clearly intended to be analogous, they *enact* a presumption that the issues share a common structure. Nonetheless, I will suggest that the form of view Shah adopts is clearer for belief and that the corresponding position on intention should be seen as an application or extension of such a view rather than a reflection of any underlying similarity.

Shah's fundamental aim in the theory of belief is to argue for a conception of belief as subject to a specific form of categorical normativity that governs belief and that is the only form of normativity to govern belief. Specifically, this is the idea of belief as subject to a truth norm: very roughly, that thinkers who accept good enough evidence for the truth of some proposition *p* are required to believe that *p*, and that if they have evidence against *p*, or if they lack enough evidence for *p* in a context in which evidence is required, they are required not to believe that *p*[24]. What represents Shah's central innovation is the connection he posits between the status of the truth norm for belief and the nature of doxastic deliberation – here understood as deliberation explicitly focused on the question of what to believe. The categorical norm thesis can be expressed in either idiom and we shall see that Shah supposes there to be a substantive relation between them. Hence Shah writes:

> one cannot settle on an answer to the question *whether to believe that p* without taking oneself to have answered the question *whether p is true*. One can certainly reflect upon one's fallibility and recognize that some of one's beliefs might be false. But so long as one is considering the deliberative question of what to believe, these two questions must be viewed as answered

[24] It is controversial how exactly to specify the truth norm for belief, and indeed the question of how to so specify it yields an informative way to think about much of what has animated the history of philosophy, particularly in relation to issues around skepticism – such as to what extent evidence for the existence of an external world is even required for such beliefs to be acceptable, let alone whether such evidence is possible. However, the precise delineation of a defensible truth norm need not concern us here. We require only the (surely defensible) idea that some sort of truth norm does indeed characterize belief.

> by, and answerable to, the same set of considerations. The seamless shift in focus from belief to truth is not a quirky feature of human psychology, but something that is demanded by the nature of first-personal doxastic deliberation.[25]

This represents the idea that doxastic deliberation is intrinsically subject to the norm that it is governed by those considerations that determine the truth of whatever is in question. Moreover, it is supposed to be governed exhaustively by this norm, so that a realization as to what the truth is automatically concludes the deliberation on what to believe. It is very important that this is not a product of the fact that a belief itself consists in an acceptance of some particular proposition as true. From that it follows only that once a deliberator realizes what the truth is, then they have acquired a belief. Shah's point is rather that once a deliberator realizes what the truth is, then they must recognize that this determines also what their belief shall and should be; it determines an answer to the question, what *to* believe. A deliberator cannot come to an opinion about the truth and yet regard that truth as irrelevant to the question of what to believe. Explaining this requires a theory of the normativity of belief and not merely of the circumstances of its psychological formation. This theory is expressed in the following passage:

> it is one thing to claim that truth has no relevance with respect to determining a certain class of beliefs (for example, about someone dear to one); it is quite another to claim that truth has no relevance to determining belief in general. I think that it is clear that we would question whether someone who made such a claim had fully grasped either the concept of belief or the concept of truth. Our handle on the concept of belief comes via its connection to truth, and this connection is an internal, normative one. Someone who claimed that truth never has a role in determining rational belief would be denying that the veracity of a belief provides even a *defeasible*, *non-overriding* reason. This, I suggest, doesn't make sense.[26]

Here a connection emerges between Shah's conception of doxastic deliberation and his conception of the normativity of belief. The latter explains the former because it is insofar as thinkers see themselves as deliberating on what to believe, and accept the idea that beliefs are subject to a truth norm, that they are thereby obliged to see the question of what to believe as answerable only with reference to the truth. This connection

[25] Shah 2003, p.447
[26] Ibid. p.454

depends on the self-consciousness of the deliberation of the thinkers of interest to Shah:

> My proposed avenue of explanation thus comes into view when we recognize that transparency [this aspect of doxastic deliberation] occurs only in the context of *asking* oneself what to *believe*. As I pointed out, this does not mean that an agent has to explicitly ask himself this question; all that is required is that the question be in the background of his reasoning, guiding his deliberation. What I suggest is that by framing his deliberation as answering to the question *whether to believe that p*, a disposition to be moved by considerations that he regards as relevant to the truth of *p* and a disposition blocking considerations that he regards as irrelevant to the truth of *p* are activated.[27]

So: according to Shah it is a conceptual fact about belief that it is subject to the standard that beliefs be correct, and it is in virtue of understanding the concept of belief that self-aware thinkers can see decisive evidence as not just typically motivating whatever beliefs they form, but also as answering the normative question of what to believe. When self-aware thinkers think about whether *p*, they are aware that they are in the business of forming a belief, and this awareness brings into play the norm governing all beliefs. Guided by this norm, such thinkers not only undergo such psychological changes as the perception of evidence happens to bring about in them, but also actively adopt those beliefs that the norm prescribes: namely, those that appear by the lights of their evidence to be true.

The process of belief formation is thus supposed to be guided concurrently by two separate elements: the first-order effect of perception of evidence, and a second-order determination to form just those beliefs that the evidence indicates would be correct. Because the second-order process involves guidance by norms in the context of awareness of the character of one's mental processes, the absence of that process would be characterized by the potential acceptance of norms that, if Shah is correct, the theorist is compelled to regard as absurd for purely conceptual reasons. Thus, a natural way to test the theory is to ask if beliefs formed in spite of Shah's suggested norms could be reasonable.

---

[27] Ibid. p.467

The theory is designed to exclude pragmatic considerations from the question of what to believe, since such considerations create nonsensical justifications for belief – justifications of the form: 'this must be the right thing to believe, because it would be nice if it were true.' Shah points out that a thinker would not be intelligible if they presented such considerations as the normative ground for their belief, unless they had independent evidence that what they found desirable to believe is also likely to be true[28].

Yet such points about the unintelligibility of certain conceptual positions are not the only proof point in favour of the theory. There is also a separate argumentative strand that concerns the immediacy of belief formation from perception of good enough evidence, most clearly expressed in the idea that there is no 'inferential step' between recognition of truth and formation of a judgment on what to believe (as already stressed, for obvious reasons there is no inferential step between recognition of truth and formation of a belief itself). It is tempting, but mistaken, to read talk of 'immediacy' in terms of psychological or temporal immediacy: as the idea that once thinkers recognize the truth, they forthwith, or straightaway, form a judgment on what to believe. Instead, the idea is that in between recognition of truth and recognition of what to believe there is no substantial ethical premise relied on. For example:

> Truth is not an optional end for first-personal doxastic deliberation, providing an *instrumental* or *extrinsic* reason that an agent may take or leave at will. Otherwise there would be an inferential step between discovering the truth with respect to *p* and determining *whether to believe that p*, involving a bridge premise that it is good (in whichever sense of good one likes, moral, prudential, aesthetic, all-things-considered, etc.) to believe the truth with respect to *p*. But there is no such gap between the two questions within the first-personal deliberative perspective; the question *whether to believe that p* seems to collapse into the question *whether p is true*.[29]

It is important that such passages disavow a psychological interpretation in favour of a logical one framed in terms of implicit intermediate premises. And when Shah asserts that 'the question whether to believe that p seems to collapse into the question whether *p* is true' this must be read in non-phenomenological terms. As we have seen, on Shah's view 'whether to

[28] Ibid. p.454-5
[29] ibid. p.447

believe that *p*' is not supposed to figure here as a question consciously pursued by the thinker; it is enough that it be in 'the background of reasoning'. So when Shah asserts that there is clearly no inferential step in between whether *p* is true and whether to believe *p*, this must be read as including within its scope just such background elements i.e. those which specifically are characterized as eluding the thinker's conscious attention. The (admittedly plausible) no-gap assertion is therefore better read as appealing to the prior commitments of the reader: it is the reader who will believe that a truth norm applies to belief and therefore that the question 'whether to believe that *p*' logically collapses into the question 'whether *p* is true', and who is likely to be buttressed in this commitment by the evident unintelligibility of the pure pragmatist's position.

These matters are important to clarify because they serve as a surprising contrast point with Shah's views on intention, views which I have already mentioned are presented as analogously structured to this theory of the normativity of belief. The theory of intention, too, is presented as a theory of the proper relationship within first-personal deliberation between a recognizable, first-order question and a reflexive, normative question: here 'whether to *A*' (where *A* is an arbitrary action) and 'whether to intend to *A*'. We should expect that the theory of intention, if it really is analogous to the theory of belief, will be structured by the same proof points and the same absurdities of rival positions. Yet this is not the case. The key passage is the following:

> When we engage in practical deliberation with an aim to arriving at an intention with respect to an action, our attention *immediately* centers on the question *whether to perform that action*. There is no inferential step between the question *whether to intend to A* and *whether to A*; the former question immediately gives way to the latter. This is why we can skip the question *whether to intend to A* and start right in with the question *whether to A* and yet be recognizably deliberating about what to intend...But if the question *whether to intend to A* were the question *whether intending to A is desirable*, there would be an inferential step between the question *whether to intend to A* and *whether to A,* bridged by the premise that the desirability of intending to *A* is determined by the desirability of *A*-ing.[30]

[30] Shah 2008, p.5

Although the assertion is identical in form to that made for the case of belief – that there is no inferential step in between the ordinary first-order question and the reflexive question – it is striking how in this discussion of intention the language is now much more psychologistic. 'Our attention immediately centers' on the first-order question: this cannot have a non-psychological interpretation. Were it analogous to belief, then it would have the structure characterized above: just as the no-gap assertion for belief really appeals to the reader's own prior commitment to the truth norm as the sole normative ground of belief, so the no-gap assertion for intention would appeal to the reader's prior belief in a good-action norm for intention: the norm that intentions are to be adopted just in case they are intentions to perform the best, or a good enough, action from among one's options, and solely in virtue of that. Yet the existence of this norm on intention, as already partially argued, is much less obvious; it is a matter for argument rather than intuition. (Subsequent chapters, especially chapter 2, substantiate this idea more fully).

This is not a minor complaint; this point about method ties in to the basic ambitions of the view. Shah's view is that, as with belief, it is part of the *concept* of intention that it is subject to a categorical good-action norm:

> According to my hypothesis, the question *whether to intend to A* gives way to the question *whether to A* because the concept of intention contains the standard of correctness: intending to *A* is correct if and only if it is not the case that one ought not to *A*. Practical deliberation, in virtue of being framed by a question that contains the concept of intention — *whether to* intend *to A* — is governed by this norm of correctness. Similarly, the question *whether to believe that p* gives way to the question *whether p* because the concept of belief contains the standard of correctness that believing that *p* is correct if and only if *p*. Doxastic deliberation, in virtue of being framed by a question that contains the concept of belief — *whether to* believe *that p*—is governed by this norm of correctness.[31]

If the good-action norm is not only true, but is part of the concept of intention, then it ought to be implicitly accepted by Shah's readers anyway, and evident in some form to them – just as the truth-norm for belief is evident. If this were true, then it would not be true that the existence of the good-action norm is a matter to be decided with reference to argument rather than intuition. Even to acknowledge that is to retreat from the claim that the existence of that norm

[31] Ibid. p.15

is evident to anyone with a good enough grasp of their concepts. And that is inconsistent with Shah's hypothesis: that the standard of correctness of intention is traceable to the concept of intention[32]. I suspect that some implicit awareness of this point is why the passage on intention veers much more towards psychological observation than does the corresponding passage on belief: it reflects an awareness of the difficulty of relying on the conceptual claim about intention in the way that conceptual claims can generally be relied on. (Since in this thesis I dispute the existence of the good-action norm on intention, I suppose I am in effect talking a subtle kind of nonsense by Shah's lights – but to the extent that my arguments even make sense, doubt is cast on this conceptual claim).

We find also that there is no analogue of the absurd fully pragmatist position on belief for the theory of intention, and this too casts doubt on the strength of the analogy drawn here between intention and belief. This is a position which we fully substantiate in chapter 3 in discussion of the Toxin Puzzle (which is a central point of reference for Shah). Here I briefly note that it is *much* less evident that an agent in the Toxin Puzzle who intends to drink the toxin is behaving absurdly; the assertion that they are is the subject of considerable argument[33]. The very fact that it is a point of argument whether it is absurd or irrational to intend to drink the toxin *itself* stands in tension with the claim that it follows from the very concept of intention, and from a norm evident within the paradigmatic use of that concept, that it is.

This is not a decisive argument; perhaps these philosophers of action really do misunderstand even what they are talking about when they talk of 'intention', or perhaps the Toxin Puzzle really does function as a device of concept clarification rather than anything else. But in order to reach that conclusion one must be satisfied that such claims as those philosophers would make are nonsensical and not merely false, and it is difficult to see how to

[32] Presumably it is a core part of that concept, in the sense that implicit knowledge of it is a condition on correct use. After all, thinking about one's own intentions is supposed to be 'governed by that norm', which seems to suggest knowledge of the norm. Given that this is so it ought to be nonsensical if intention is spoken of in a context in which it is doubted. Shah is free to argue that this condition is buried more deeply in the concept than is evident to its users, but I leave this potential line of argument to him.

[33] E.g. Gauthier, 'Intention and Deliberation', in Danielson (ed.) 1998.

reach that conclusion to the extent that one can read and understand that sort of literature and coherently entertain what it would be like to be the sort of agent whose intentions are formed as these philosophers suggest.

There is a further, specific reason why it is difficult to render an analogue within the theory of intention of the absurd fully pragmatist position for belief. Suppose one is asking whether it makes sense to adopt an intention to raise one's arm now because (say) a scientist with an interest in intention-relevant neurology would like to take a scan of someone in possession of a present-directed intention (i.e. an intention to raise one's arm in the present) and they will offer one a reward for having such an intention. The way to get that reward is simply to raise one's arm, since, as Shah points out, to do that is a way of ensuring that one does have that intention[34]. As mentioned earlier, such actions are a *sine qua non* of desirable intentions, so there is a reason to perform them. This only complicates the issue; it means that the Toxin Puzzle, where the rewarded intention in question is future-directed, is the only way one could in principle construct a case where a pragmatist view of intention would yield a different verdict from Shah's position.

Is Shah's idea that in practical deliberation our 'attention immediately centers' on the question 'what to do' itself plausible, construed as a purely psychological observation? It is – but now there is trouble in connecting this observation to the theory of the normativity of intention that Shah wishes to adopt. The question 'what to intend?' is not itself explicitly asked in normal practical deliberation, so the point that our attention immediately centers on the question 'what to do' does not prove that there is any kind of logical or conceptually derived collapse.

Moreover, unlike the case of belief, there is no sense in which arriving at a judgment on what to do is *itself* to arrive at an intention – or at least, this is the position I shall assume: that normative judgment and intention are distinct (fuller discussion of this issue takes place in chapter 5). Were Shah's theory of intention to be correct, then deliberation on what to intend, insofar as that is entirely determined by reflection on what to do, could *only* be driven by an

[34] Shah, ibid., p.5

awareness of the putative conceptually based norm that intentions are to be directed at the better actions from among one's options, i.e. based on one's normative judgments about action. It is difficult on this view to explain how practical deliberation issuing in an intention could work prior to the acquisition of (and acquiescence in[35]) the relevant concepts[36]. If one rejects Shah's position, or even if one thinks it is correct while also thinking that practical deliberation requires more than adherence to these conceptual norms expressed implicitly in judgments about what *to* intend, then there is an *additional* explanandum created here. Namely, one must explain *how normative judgments influence intentions* without appeal to the reflexive norm awareness Shah posits. This is a genuine problem and one for which Shah's putative solution should count as a central point of reference.

This new explanandum emerges only when we step back from the analogy with belief. In the theory of belief the issue is straightforward since belief is identical to an acceptance-as-true of some relevant proposition. Since intention, I assume, is not identical to a normative judgment[37], it is to be explained how normative judgments are capable of influencing intentions. On *this* point the psychological observations seem much more relevant since in any conscious deliberation an intention will be an upshot if the agent reaches a normative judgment and then, as they usually do, adopt an intention on the basis of that. The point that our attention 'immediately centers' on the question what to do, and that the consideration of this question is usually sufficient to determine intention, itself demands explanation; that is, not only the influence of normative judgment on intention, but the immediacy of that

[35] If the determination of intentions by normative judgments on action were solely based in the concept of intention, one could ask what grounds there were for adopting just this concept of intention rather than some different concept. That agents could do better if they were able to pragmatically regulate their intentions would be evidence in favour of adopting a different concept of intention, one not subject to the good-action norm. This point is not quite in tune with the ambitions of Shah's view, on which the good-action norm governing practical deliberation is categorical – not up to us to change. It is interesting to consider to what extent a similar objection might also be made against Shah's conception of belief.

[36] One test case for this issue would be the possibility of intention in non-ratiocinative and pre-ratiocinative creatures – a point emphasized by Shah 2003, p.468. If such creatures turned out to be incapable of intention, this would be a point in favour of Shah's view: it would suggest that the capacity to adopt intentions requires conceptual capacities, and one way to explain how this might be so is if there is a constitutive link here.

[37] This is also Shah's position; cf. 2008, p.3.

influence in normal conditions, demands explanation, along with the absence of any (conscious) inferential step here. I address this question in chapter 6.

As a psychological observation, however, it is inconclusive, and refuted by appropriately plausible and vivid counterexamples of the rational determination of intention without reference to, or in spite of, the agent's normative judgments about action. Counterexamples like this are supplied in chapter 2. But any failure of generality here should not render inert the power of the intuition Shah supplies: that often no inferential step is needed between normative judgment and intention in order for intention to be rationally determined by that judgment: from the perspective of first-personal practical deliberation, there is often some kind of collapse, and this deserves to be explained in terms more serious than that the relevant inferences are in fact unconscious. The intuition in effect says that a collapse within practical deliberation makes sense conceptually and not just psychologically; on this point Shah is correct. Reconciling the legitimacy of a conceptual collapse in some cases with the possibility of rational pragmatic determination of intentions in other cases takes some work, and this is attempted in the following chapters.

So: the unobviousness of the good-action norm for intentions puts the comparison with belief in difficulty. The hypothesis that a conceptual norm governs practical deliberation, directing us to form intentions in accordance with our normative judgments about action, cannot be as easily substantiated as the corresponding position on belief. There is no particular reason here to assent to the good-action norm or the conceptual hypothesis that would underpin its existence. However, there is a valid point about immediacy, or the absence of an inferential step, for practical deliberation, which will be addressed.

Intention and Belief 2: Hieronymi on intending at will and believing at will

The previous section discussed the idea that good action stands to intention in something like the relation truth stands to belief: as what the thinker/agent must take to be the case in order for them to rationally adopt and maintain

that belief or intention[38]. The next section explores a separate motivation for excluding state-given reasons from the justification of intention: that the same grounds which lead us to affirm that belief is a non-voluntary response to the evidence should lead us to affirm that intention is a non-voluntary response to reasons to act. I shall start by explaining the situation for belief, before then asking whether it can be successfully carried over to the case of intention. This argument has been pursued by Pamela Hieronymi in 'Controlling Attitudes' (2006).

Voluntariness is relevant here because the issue of whether or not belief is voluntary connects, in multiple ways, to an underlying conception of belief as necessarily responsive only to reasons bearing on the truth of what is to be believed (*mutatis mutandis* for intention). The non-voluntariness of belief is treated as evidence for this conception because the responsiveness of belief to evidence bearing on the truth of its content would, firstly, be a good explanation of why one cannot wilfully choose to believe any arbitrary proposition – and secondly, a proof for the non-voluntariness of belief. Williams writes:

> One reason [I cannot believe at will] is connected with the characteristic of beliefs that they aim at truth. If I could acquire a belief at will, 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.[39]

This passage displays both argumentative tendencies. If we were already convinced that thinkers can't will beliefs into existence, then the point that that power of choice would imply the ability to choose beliefs whether or not they're true might explain why. Knowing that they believe because they choose to believe, the thinker has no grounds to suppose that their belief represents the truth. But since to believe is to think true, the thinker has no grounds to suppose that what they take to be true is true. Knowing this, it is unclear how they could continue to believe it. If this is a good explanation, then it makes attractive the underlying conception of belief. On the other hand, if we are convinced that beliefs 'purport to represent reality', then this

[38] Cf. Tenenbaum 2020, p.5-6 for more discussion.
[39] Williams, "Deciding to Believe", in Williams 1973.

same consideration would lead us to object to the idea that believers can ever choose what they believe.

But is this a good explanation, or a good argument, whichever direction it runs in? The dialectic must go deeper than the brief outline suggested above. Simply knowing that one lacks justification for one's beliefs isn't automatically disqualifying – or at least this itself requires further argument. For one thing, foundationalist positions assert that there are some beliefs that are not based on anything further; beliefs that don't have any vindicating explanation[40]. The point here is not that we acquire those beliefs indifferently to whether or not they are true, but rather that we lack further, independent evidence to support them. Secondly, even setting aside these classical philosophical questions, even some apparently basic beliefs might not be based on anything like grounds (such as the belief that one is in pain). While such beliefs are dubiously able to be acquired at will, arguments about grounds for belief don't capture why. The important difference is what the difference is here between such cases and what is supposed by the voluntarist to be possible: if it is ever acceptable for a thinker to maintain a belief while lacking evidence to support their belief, why should we suppose that it is never acceptable for beliefs acquired at will to lack independent evidence?

However, there is a limit to how compelling this particular comparison can be. The beliefs which are supposed to be foundational beliefs have *many* special features that are lacked by the more mundane beliefs which the voluntarist and their critic are typically concerned with. Foundational beliefs may be the foundation of a whole train of beliefs and inquiries that, collectively, result in an increased ability to make sense of the world; they may be psychologically inextinguishable; they may be irrefutable. Such characteristics may well not be possessed by many of the belief which the voluntarist would suggest could be acquired at will.

A more significant criticism of this version of Williams' anti-voluntarist argument is that the relationship between belief and credence seems to permit

[40] For example, Hume (2009), 'On scepticism with regard to the senses', on the belief in object permanency and externality. Even if such beliefs are necessary, their psychological origin does not constitute an indication that such beliefs are even meaningful, let alone true.

another potential entry point for non-evidential impacts on belief formation. It is sometimes thought that whether it is permissible to believe *p* depends on how much credence it is rational to have in *p* combined with a threshold, determined in context, for how much credence is necessary for belief to be permissible[41]. The threshold for acceptable belief is not itself a function of the evidence for *p* (that is, rather, in the domain that determines acceptable credence). Much more argument would be needed to establish that evidential thresholds for belief could be reasonably subject to the will. But Williams' argument, at least, does not rule out the possibility that they could. It does not rule out this way in which, on occasion, practical and pragmatic justifications could be supplied for whether to adopt a particular belief, through the voluntary adjustment of the relevant thresholds. (Of course Williams is not denying the influence of the will on belief, but rather is talking about whether a belief formed in that way could ever be justified).

For all that has been said so far, and that is suggested by the brief argument given above, is that the only thing that would be doxastically disqualifying is if one's beliefs were contrary to the balance of evidence; yet being contrary to the balance of evidence is not an implication of a belief's having been chosen. The fact that beliefs 'purport to represent reality' seems to show only that one could not coherently choose to believe what one takes to be against the balance of evidence. Yet Hieronymi claims that *no* belief can be chosen, irrespective of its evidential status[42]. This claim is significantly stronger.

It seems that what is repugnant about willed beliefs is not that the thinker lacks justification for their truth but something more. In particular there seems to be more significance to the fact that the thinker is indifferent to the truth when they will the belief into existence: they decide to believe that *p* not just in the context of uncertainty about *p* (for all their evidence suggests) but rather while suspending a concern for *p*'s truth in the context of their deliberation on whether to believe it.

---

[41] Cf. Jackson 2018, section 5 for an examination of this issue.

[42] 'I hope to show that believing could not be "voluntary," that is, one could not believe in the way one can perform an ordinary intentional action.' Hieronymi 2006, p.45.

This connects more directly with Williams' idea that beliefs *aim* at truth. The picture this suggests is as of a practical project in which the believer tries to obtain true beliefs: the believer cares about what is true and tries to find methods to adjust their beliefs accordingly. In contrast, the thinker who wills a belief into existence, no matter how much doing so is rewarded, 'irrespective of its [the belief's] truth' is suspending any such concern with the truth. On this reading, what the passage argues is not strictly that beliefs cannot be willed into existence, but rather that the only context in which it makes sense to adopt a belief that *p* is one on which one is concerned to represent whether *p* according to whether or not *p* and adopts (or not) a belief that *p* as part of that project. This idea introduces, at a minimum, a very substantial restriction on the contexts in which willed belief could make sense. To will a belief arbitrarily (just because one feels like it) or to will a belief for the sake of financial or divine reward would involve dispensing with a concern for truth; such contexts are among those in which the thinker cannot conceive of what they are doing as 'purporting to represent reality'.

This discussion has set up the dialectic with respect to belief with a view to examining to what extent similar sorts of arguments can be constructed for intention. Our question is: can any argument for the responsiveness of intention to reasons to act alone be constructed on lines analogous to the above? Hieronymi argues so. Noting that it is 'quite standard'[43] to think of intention as settling the question what one will do, she suggests that intention is constituted by 'a commitment to doing something': it involves a commitment to performing the associated action.

Then, the second premise. 'A reason is a consideration that bears on a question…the reasons will already specify the question under consideration'[44]: to consider any reason for an action is to consider whether to perform it. Finally, answering a question *is* acquiring the relevant commitment: 'when one answers a question for oneself (again, however

[43] Ibid., section 3, p.56
[44] Ibid., section 5, p.59

implicitly or unreflectively), one might therein, *ipso facto*, arrive at a belief or an intention'.

This leads Hieronymi to a very general argument, one that carries through the fundamental conception we are exploring of the similarity of intention and belief. 'Since the reasons will already specify the question under consideration, and since the question determines which attitude will be immediately formed or modified… the agent does not… have discretion over which attitude is controlled in response to which reasons. Rather, in taking a consideration to be a reason, she has already determined which attitude she will evaluatively control in response to it.' Thus, the agent who considers the reasons whether to φ, if those reasons are decisive and correctly taken up by them, will reach an answer to the question whether to φ, and thereby acquire an intention-constituting commitment to φ.

Questions regarding voluntarism pertain to these issues because the very thing that, according to Hieronymi, explains why belief cannot be adopted at will – the fact that belief *aims* at truth and so involves a constitutive commitment on the truth of what is believed – explains why intention can't be adopted at will either: intention involves a constitutive commitment on whether to perform the intended action. This is offered in the case of belief as the best explanation of nonvoluntarism about belief, and then generalized to include intention, so that we are to conclude that intention, too, cannot be adopted at will.

An initial response to this argument is to repeat a point made above: that intention involves settling what to do does not entail that the formation of intention is subject *only* to questions of whether the action in question is worthwhile. But this does not exhaust the suggestiveness of the comparison with belief in respect of voluntarism, given the different possible readings of that argument in the case of belief examined above. I shall argue instead that there is not a sound comparison here.

If there were a strong comparison to be made between belief and intention on this issue, we would expect that we would ultimately be able to translate the Williams passage that argues against belief voluntarism into language appropriate to the case of intention. But different possibilities present

themselves as the appropriate focal point for the analogy: if belief aims at truth, what does intention aim at? One possibility is that agents aim at doing the right, or best, or good enough, or appropriate, thing when they act. That is, after all, the thesis of our opponent: that rational agents conform their intentions according only to the characteristics of the action it would be an intention to do. Reconfigured in this way, the Williams passage would read:

> One reason [I cannot intend at will] is connected with the characteristic of intentions to act that they aim at appropriate action. If I could acquire an intention at will, I could acquire it whether or not its associated action was useful, proper or good. If in full consciousness I could will to acquire an "intention" irrespective of the characteristics of the action that would be intended, it is unclear that before the event I could seriously think of it as an intention, i.e., as something purporting to represent what it is appropriate to do.

The idea here is that agents can only conceive of themselves as properly *intending* to act when they aim at doing the appropriate thing. The difficulty is that this idea is obviously false. It is false firstly in some cases of akrasia or weakness of will, where one knowingly intends something against one's better judgment as to what one should do. Though akrasia obviously involves some irrationality, it is not the sort of irrationality that disrupts an ability to think of oneself as intending. In contrast, if one were epistemically akratic, then to know that would compromise one's ability to think of oneself as believing. For example, if you knew that you believed that someone's action was vile only because it allows you to criticize them, then you wouldn't see yourself as really *believing* it – rather, you would see yourself as merely being prepared to perform that narrative[45].

Secondly, it is possible to form intentions to perform utterly indifferent acts. An agent can enter their bedroom and quickly, intentionally, look right and then left before moving on. Such an act is completely neutral in respect of the reasons. There is no particular reason to keep one's eyes focused on a

[45] There is some room for debate here: some cases perhaps approximate to epistemic akrasia and it is debatable whether they exemplify it. For instance, suppose you really want to give someone the benefit of the doubt because it would be inconvenient to have to confront them about what they may have done – and consequently, you believe that they are innocent despite moderately good evidence to the contrary. Is this epistemic akrasia? It would be only if the evidence was compelling; if the evidence itself leaves room for doubt, then it would seem to resemble motivated suspension of judgment, or a motivated raising of the threshold for convincing evidence. This is not quite analogous to the practical case.

particular direction when entering one's bedroom (usually). There are all kinds of trivial acts which an agent may intend to do. In intending to do them, they are not conceiving of themselves as doing the *appropriate* thing, but this in no way disrupts a possible conception of themselves as intending to do it.

Since it is possible to think of one's intentions as failing to represent what it is appropriate to do, it is possible (for all this argument suggests) to acquire them at will irrespective of whether the associated action is appropriate. Doing so would not hinder our thinking of them as intentions. *If* there is a constraint of this sort on intentions, one that precludes their being based on anything other than reasons to act, then it would have to be a constraint of a different kind (one that does not hinder the description of intentions violating that as genuine intentions) and with a quite different underlying basis (e.g. some kind of empirical, purely psychological impossibility).

The idea that beliefs aim at truth, that beliefs only make sense in the context of attempting to represent how things are, found its analogy in the idea that intentions aim at appropriate action and only make sense in the context of trying to act well and forming one's intentions by those lights. I suggest in contrast that intentions make sense in a broader range of contexts and so cannot be thought to have that constitutive aim – a claim more fully substantiated in the next chapter. But this leaves intact the weaker objection which we put forward initially in the case of belief: the idea that adopting beliefs at will would dissolve any necessary connection between the thinker and any grounds for supposing their belief to be true, which would itself compromise the belief. The equivalent of this objection, now for the case of intention, is that the ability to adopt intentions at will would dissolve any necessary connection between the agent and any grounds for supposing their intention to be appropriate. Does this objection make sense?

Not quite. Again, there is a multiplicity of options for filling out the objection. One way of filling out the objection is to say that adopting intentions at will dissolves any necessary connection between the agent and reasons to have their intention. But this is not true. Intentions behave like actions in this respect: intentions that can be adopted at will will be adopted when the agent

has reason to adopt them – or else adopts them arbitrarily, in just the same way that actions are either governed by practical reason or else done arbitrarily.

The real equivalent version of the objection, the one preferable to our opponent, is that the ability to adopt intentions at will means that it is possible for agents to adopt intentions irrespective of whether the intended action would be appropriate. In the case of belief, it seemed that – setting aside exceptional cases – without a further justification in support their belief, it is unclear how the thinker could hold onto that belief. They would, rather, be obliged to be uncertain; they cannot rationally suspend the project of checking their beliefs for truth in accordance with the evidence available, and when they did check their willed belief, then, knowing they lack justification to believe it, they would update accordingly into a state of uncertainty. For intention, the analogous thought would be that without reason to do what it is an intention to do, it is unclear how the agent could maintain the intention. They cannot rationally suspend the need to update their intentions in accordance with their reasons, and if they decide that they lack reason to do what it is an intention to do, then why would they continue to intend to do it?

But this way of putting the objection fails at the last step. What we are supposing is that an agent may maintain an intention to act, because it's a good or useful intention to have, even if the associated action is pointless in itself. If an agent decides that they lack reason to do what it is an intention to do, then, in such a case, they still have reason to intend to do it: whatever reason is provided by the usefulness of the intention itself.

We can confirm this further by trying a second substitution-pattern for the Williams passage, one on which intention aim only at the performance of action:

> One reason [I cannot intend at will] is connected with the characteristic of intentions to act that they aim at the performance of an action. If I could acquire an intention at will, I could acquire it whether or not I was going to perform its associated action. If in full consciousness I could will to acquire an "intention" irrespective of whether I would perform its associated action, it is unclear that before

> the event I could seriously think of it as an intention, i.e., as something purporting to represent what I shall do/what to do.

While it is plausible that if I don't represent some action as what I shall do, then I cannot think of what I am doing as intending it, this in no way casts doubt on the possibility of my acquiring such an intention at will. For me to acquire such an intention at will would also be for me to aim at the performance of that action, something which is directly up to me in the same way that the formation of intention is. My intention will then represent 'what to do' – it will represent the action at which I am aiming. This says nothing about the necessary grounds on which such an intention must be adopted. Though there may be constraints on when exactly it is possible for me to aim at the performance of some action (for instance, I may be unable to do so if I know that I can't succeed) those will also be constraints on when it is possible for me to wilfully adopt an intention.

These points are blurred in Hieronymi's discussion because of an ulterior issue about the way in which she characterizes reasons to act. If any reason that helps to settle the question of what to do is a reason to act, then since to have an intention is to be committed on what to do, any reason that helps to settle what intention to form helps to settle what to do. This obscures the distinction between reasons to act and reasons to intend, or in other terms we have used, between object-related reasons for intention and state-given reasons for intention. It trivializes the claim that intentions are determined only by the agent's perception of what reasons there are to act, since reasons to intend must be a species of reasons to act, on this view, in helping to settle the question of what to do through settling what to intend. That is, it lends false credibility to the claim that intentions are determined only by the agent's perception of what reasons there are to act, an idea that is then treated as equivalent to the more substantive, non-trivial claim that state-given reasons for intention aren't legitimate. To recover the question of whether one might intelligibly intend at will for state-given reasons, we must set aside this characterization of reasons to act.

So both versions of the non-voluntarist objection to the legitimacy of state-given reasons for intention fail. This in turn tells us something about the

differences between intention and belief. Though it is plausible that belief must be non-voluntaristic and must purport to represent reality, as Hieronymi and Williams suggest, there is no reason here to think that something similar is true of intention.

Conclusion

This section has so far pointed out that intention works differently from belief and cannot be connected with the need for justification that bears on the content of the intention in the same way as it was obvious for the case of belief that it generally requires justifications bearing on the truth of what is to be believed. But to confirm our interpretation of the differences here we need to supply a broader explanation of why they hold. This project, of charting differences between theoretical and practical reason, will recur throughout this thesis. What is needed is an account of why believing only makes sense in the context of aiming at the truth, whereas no corresponding constraint holds for intentions in respect of actions – or at least not a constraint of the same kind, that can be argued for in the same way. These questions are addressed further in chapters 4-6.

The next chapter seeks to substantiate further the claim I have made in this chapter that there is no good-action norm for rational intention: it provides a body of cases to substantiate this. Now that the ground has been cleared for thinking that intention does not have to be conceived along lines similar to belief, the way is now open to entertaining the idea that it differs also in terms of its relation to its objects. The next chapter in particular argues that relatively mundane cases allow us to see the legitimacy, and even the indispensability, of state-given reasons for intention.

2. First-order evidence for state-given reasons for intention

This chapter discusses attempts to prove via example that it is sometimes rational to form or not form intentions because of the benefits of having or not having those intentions rather than the worthwhileness of what they are intentions to do. This chapter attempts to prove a strong version of that thesis: that state-given reasons, even those that derive from highly extrinsic benefits, are rationally admissible to questions about whether to form even particular intentions to do or not do particular things. This question is tackled here via a series of discussions of particular cases: examples where it is apparently rational to form particular intentions to do particular things. If these examples are genuine, then the thesis must be accepted: state-given reasons at least sometimes make a difference to what intention it is rational to adopt. Moreover, if the thesis is true, it would be extremely surprising it could not be proved via illustration; we would expect that there would be cases in which state-given reasons would make the decisive difference to which intention it is rational to form. It is therefore important to devote space to discussion of relevant cases just so they can be looked at in their own right.

An important point of disagreement among those who are favourable to the admissibility of state-given reasons for intention is what sort of state-given reasons are rationally admissible. Weaker views will be examined first; the view put forward here is that there are very few or no restrictions, and that will be argued for at the end.

Schroeder on state-given reasons

Schroeder has suggested that only very limited sorts of state-given reasons for intention are rationally admissible; surprisingly, he thinks also that this

sort of state-given reason also applies to doxastic deliberation. These are reasons specifically *against* forming particular intentions/beliefs[46].

He gives three examples of such reasons: firstly, 'If the evidence is too evenly balanced or merely probabilistic in nature, then that can make it irrational to make up one's mind'[47]: the fact that the evidence is finely balanced is a reason against forming an intention/belief either way. Secondly, the fact that decisively important information is forthcoming, in particular, information that bears on the value of an option. If the agent soon expects to learn whether or not option A is preferable to option B, then that counts against now having an unconditional intention to take either A or B. Schroeder illustrates this idea with the following case: he does not know whether his brother will be in town tomorrow; if he's there, it's worthwhile going into town, but not otherwise. His brother will call later today to say whether or not he'll be there. The suggestion is that:

> waiting to decide is the only rational course. It is not only rational *for* me to form neither intention now; if I do take the fact that my brother will call me later this afternoon into account and wait to decide for that reason, then I am rational *in* forming neither intention now. And there is no intuitive difficulty in forming neither intention on the grounds that more information will soon come to light—on the contrary, it is easy to wait to decide for exactly this sort of reason.[48]

The third example stems from the co-ordinatory benefits of intentions: it can sometimes be worthwhile not to change one's intentions if other people are relying on you to persist in a particular intention. So the agent, in such cases, has reason against adopting any other intention than the one they have so far – so it does not depend on whether the other intentions would be intentions to do more worthwhile things than what one currently intends. For example, if the agent has already decided not to go into town, then his wife may rely on unfettered access to the car; if he changes his mind about whether to go into town, this will spoil or at least complicate her plans, which is a reason against changing his mind.

---

[46] Cf. Schroeder 2013, p.131-2: 'my basic view is that…it should simply be immediately deeply puzzling why anyone would ever think that the object-given/state-given theory might be true of reasons *against* belief and intention, as well as reasons in favour.'
[47] Schroeder 2012, p.478
[48] Ibid., p.467

The inclusion of these sorts of reasons in the general category of 'state-given reasons' appears puzzling, since these sorts of reasons clearly have much to do with the object of the intention or belief and engage fully with what the formal aims of each state are supposed to be: truth in the case of belief, or good action in the case of intention[49]. A believer who aims at truth will obviously be concerned with whether or not their evidence is too finely balanced to allow them to make a determination, since lunging one way or the other may well lead them into error; an agent who aims at acting well will also be concerned with whether they are in a sufficiently informed position to legitimately make up their mind, since they can potentially improve their chances of acting well if they wait until all the relevant information is in. This is because the agent needs to make up their mind at the right time. If they make up their mind too early, then later information may not be very effective; it would require the agent to re-open the issue, and the agent may not be willing to do that if they have already made other plans. It is therefore in the agent's interest if they postpone making up their mind in order to be able to make a more informed determination on what would be the best thing to do.

Although these two first two kinds of reasons can be accounted for in terms of the idea of the formal aim of intention, the exception here is the case of co-ordinatory benefit, but that example is not very convincing: if the agent has reason not to change their plans not to use the car because their wife plans on using the car, then they do indeed have reason not to change their plans, but here they also have reason not to use the car (given that plans for its use have already been made), since if they use the car, that will then deprive the wife of the opportunity to carry out the plans she has made. So it is not clear in this example that the reason not to change one's intentions does not correspond to a reason not to perform the action that whose non-performance the wife is relying on. If the agent has reason not to use the car, and also a reason not to

[49] The idea of good action as something the agent necessarily aims at is more often propounded with respect to action itself, so that in action, the agent necessarily aims at acting well (cf. Tenenbaum, The Guise of the Good for an overview of such positions). Extending it to intention is an extrapolation, but one whose usefulness I hope is apparent in this context, and that certainly compares to the positions of Shah and Hieronymi examined in the last chapter.

plan to use the car, then it is not a case in which there is a clear state-given reason for intention that goes beyond the agent's reasons for action. The argument would have to be that the state-given reason is not derivative on the object-related reason, but that would take more subtle argument that Schroeder does not supply.

The first two cases are thus more relevant to the argument, and they are the ones in which it seems that the reasons do distinctively bear on the notional formal aim of intention, namely good action. This relevance to what the intender is supposed to aim at (according to the theory we reject) makes those reasons distinctively object-related in a sense not shared by the more obvious examples of state-given reasons. This is a point Schroeder himself makes: he contrasts his preferred examples of acceptable state-given reasons with more generic state-given reasons like being offered money to intend one way or the other, suggesting that the latter kind of reason does not suffice to make intention rational, whereas the former does[50].

Instead, Schroeder defines the object-given/state-given distinction as follows: if (as on some theories) only object-given reasons are rationally admissible to deliberation about what to intend, then any reason R 'is a right-kind reason [i.e. an object-related reason] bearing on intending to do A just in case R is a reason bearing on whether to do A.'[51] And Schroeder specifies a crucial further part of this way of defining object-related reasons, that 'bearing on whether to do A', here, really means in favour of doing A or in favour of doing not-A. He writes:

> So long as the reasons bearing on whether to do A are exhausted by the reasons to do A and the reasons to do not-A, it is a consequence of [this condition] that R is a right-kind reason bearing on the intention to do A just in case R is either a reason to intend to do A or a reason to intend to do not-A. Hence, [this condition rules out] right-kind reasons to lack intention.[52]

Schroeder's argument is that since reasons simply to lack or not to change intention/belief don't satisfy this condition, because they neither justify the performance nor the non-performance of A, they are state-given reasons, yet

[50] Ibid., p.469
[51] Ibid., p.464
[52] Ibid., p.470

rationally admissible, or 'right-kind' ones. A reason to lack intention, on Schroeder's conception of object-related reasons, would have to be a reason to not perform the action it is an intention to do.

Even if the examples are successful[53], this aspect of Schroeder's preferred state-given reasons fails to connect with part of the thesis put forward here: that it is *because* intention lacks a formal aim that state-given reasons are generically appropriate and admissible to questions of what to intend. If intention does possess a formal aim (and if bearing on the formal aim defines object-relatedness), then, as already argued, Schroeder's examples, even on his preferred interpretation, are not convincing. An agent who aims at good action may sensibly refrain from intentions when important information is forthcoming, waiting until then to make up their mind. Doing so contributes to their aim of acting well. Exactly this defines the sense of object-relatedness that is the focus of this thesis; accepting Schroeder's versions of state-given reasons would not help to prove the ideas propounded here.

Pink on reasons to decide

A more useful, extended case is provided by Pink[54]. To clarify how it works, it is first useful to set up the general sort of state-given reason which the case is supposed to illustrate.

It stems from a background conception of intention as subject to a formal aim that is similar to, but not the same as, the one we have centrally been discussing so far: the idea that intentions aim at worthwhile action. On Pink's view, intention does not aim at worthwhile action, but rather aims at *producing* worthwhile action – a doctrine Pink terms 'the practical primacy of action':

> Decision-making, I have supposed, is an executive, reason-applying agency. Its function is to help apply practical reason as it concerns action. So when practical reason recommends that we take decisions, this is characteristically

[53] Cf. Shah and Silverstein 2013, Hieronymi 2013, Hubbs 2013 for discussion; Schroeder 2013 for a reply.
[54] The case is the central point of discussion in 'In Defence of the Action Model', chapter 8 of Pink 1996.

> because so doing will help us apply practical reason's recommendations of *actions*… [The function of decision-making] is to ensure that the actions which we perform at any given time are those which are justified given the actions which we perform at other times…There is a *practical primacy of action.*[55]

It is clear that this is highly similar to the good-action norm examined in Chapter 1, but the emphasis on *applying* reason and *producing* actions modifies the core idea and allows an additional sort of state-given reason to count as legitimate – legitimate because it bears on an intention's producing worthwhile action. If intention is to produce worthwhile action, then it follows just from that characterization that intention is not usefully adopted where it cannot fulfil that role of producing the action intended[56]. When some psychological property of an intention bears on its ability to produce worthwhile actions, then it is relevant to this formal aim and so can become a state-given reason for or against certain intentions. (For our purposes, it is sufficient for a state-given reason that it does not reduce to a consideration one can bring to bear on the actions themselves – thus the reason that stems from the psychological property of the intention here counts as a state-given reason by our lights, despite the fact that it is legitimated in virtue of the putative formal aim of intention).

Pink illustrates this with a case in which an agent ("Dan") must take into account his own likely future preference shifts because if they occur Dan is somewhat likely to change his mind on what to do, and he is then likely to act in a suboptimal way. This makes it sensible for Dan to adopt a plan that delivers a good enough outcome whether or not those preference shifts occur: Dan must reject plans that he cannot rely on carrying out. That Dan is somewhat likely to abandon a plan midway and crash out into a bad outcome gives Dan a reason not to make that plan. The unreliability of the intention makes for a state-given reason not to have it, in a way that makes sense in light of the formal aim of intention/decision. This is the essence of the case.

---

[55] Ibid., p.209

[56] Pink makes a similar point in *Self-Determination*: 'Deciding to attain E is only justified if taking that decision is likely enough to bring E about. Which is why sensible people don't take decisions about matters their decisions clearly can't affect; since the function of decisions is to lead to their fulfilment, that a decision has no chance of doing this is a conclusive argument against taking it.' Pink 2016,, ch. 11 section 2, p.199

More specifically, the action Dan considers, and potentially expects to prefer later, is to perform a stunt. Performing the stunt is something it is better for Dan to do in the context of various preparations (organized publicity for the stunt) designed to enable Dan to fully exploit the potential benefits of that action (presumably glory, or money). Yet those preparations themselves ought to be made only if Dan goes on to perform the stunt – otherwise there will be costs. So the best outcomes for Dan are: to organize publicity and to perform the stunt, or else to do neither. Because Dan is somewhat likely to end up wanting to perform the stunt no matter what prior plans he has made, there is a risk that if he doesn't organize publicity he will perform the stunt *anyway*, and this would be bad – worse than doing nothing at all (that is, performing the stunt without publicity is worse than neither organizing publicity nor performing the stunt). It would be better if the stunt were performed with publicity already organized.

So, because an intention to do nothing (neither organize publicity nor perform the stunt) is unreliable, this creates a reason for Dan not to have that intention that is not a reason against the object of that intention. That is, the fact that Dan is likely to want to perform the stunt anyway does not tell against the value of doing nothing (of neither organizing publicity nor performing the stunt). It is not a reason against this combination of (in)actions. It is just a reason against *planning on* that combination, because a plan is inappropriate if one cannot rely on oneself to carry it out, and if there is an alternative plan that delivers just as good an outcome that one can rely on oneself carrying out. It is this precise fact that makes it a state-given reason.

This is a plausibly genuine state-given reason, whether or not we accept the suggested aim of intention as a restriction on state-given reasons. In discussing it, we can immediately make a similar point that we made in discussing Schroeder: that it simply does not help to prove the thesis suggested here, because it only presents a very limited sort of state-given reason, namely that bearing on the efficacy of intentions, whereas this thesis is devoted to state-given reasons much more generally. Before discussing some of Pink's arguments for why state-given reasons must be limited in this

way, we can note a general implication of the veracity of this kind of state-given reason.

The point that intentions play a productive role in relation to action ought to allow simpler examples of state-given reasons using just the supposition that decisions to act, if taken, make the eventual attempt of that action more likely than if that decision had not been taken. For example, consider a case in which someone thinks it is best all-things-considered that they confront their boss at work, but also knows that it is significantly likely that they will chicken out or procrastinate over this confrontation. It is not difficult to imagine, just intuitively, that they are more likely to go through with it if they very purposefully make a decision at the beginning of the day that today is the day that they'll finally confront their boss. If this is right, this provides a strong state-given reason for taking that decision: it will help them do what it is best that they do. Here, as in Pink's case, it is the psychological effect of the decision that provides the state-given reason. This is not a reason for them to make up their mind one way or the other; it is a reason specifically for them to make up their mind one particular way.

There is empirical support for the idea that adopting an intention to perform some action does significantly alter the psychological relation in which the agent stands to that action in relevant ways, particularly in ways that may conceivably generate positive state-given reasons to make decisions. A philosophical presentation of the relevant evidence is given in Holton (2009)[57]. Agents with intentions show certain psychological changes relative to agents without intentions that go beyond the basic collection of attitudes involved in having a plan or aim (it is suggested that the forming of the intention is responsible for these changes). In particular, agents tend, among other things, to become much more confident in their abilities to do what they intend, even to the point of overestimating the extent to which they can exert control over their situation[58]. They are also more likely to act on their

[57] Holton 2009, chapter 1, p.5-9.
[58] The overestimation occurs even with respect to situations unrelated to the intention itself; it extends even to ideas of how much their actions affect visible aspects of their environment. Holton (ibid.) presents one experimental paradigm in which it was found that

intentions when they form very specific intentions instead of general intentions whose details have yet to be worked out, even to the extent that subliminally presented cues can trigger the relevant behaviour, provided that the specific intention has already been formed and the goal has not been abandoned.

These particular two changes are remarkable here because it is obvious how they can make it useful for the agent to deliberately adopt certain intentions, or to adopt specific intentions over general intentions. They make plausible the situation suggested earlier, in which an agent is well-advised to consciously make a decision at the beginning of the day to confront their boss on that day, instead of leaving it until later to judge whether today is the right day; another example would be a would-be public performer who is nervous about their capacity to perform, but who is well-advised to decide to do it anyway, in order for them to reap the likely benefits of confidence over-estimation. It would be surprising if the only way agents could get themselves to adopt such intentions is through indirect or non-rational measures that would somehow induce them to make the right decisions. Rather, it would be rational for them simply to decide to do those things, for those reasons. And that suggests that these state-given reasons are rationally admissible to their deliberation on what to intend.

Special purpose agency

Points about the psychological power of decisions, or of the mindset into which the resulting intentions put us, make plausible the basic possibility of state-given reasons for decisions and intentions. They do not decisively establish that possibility because it is open to a sufficiently determined theorist to argue that, though such factors may establish the *usefulness* of such decisions and intentions, they don't establish that it is rational to actually form those intentions[59]. This position itself has intuitive evidence against it: it is

subjects who focused on their goals were much more prone to overestimate how much effect a button-pressing of theirs had on turning on a light.

[59] Cf. Hieronymi 2006, p.57 for this response to such cases.

intuitive that it is rational for those agents to decide, in the normal way decisions are made, to perform the relevant actions.

Apart from the blanket denial of the rational admissibility of state-given reasons that the pure good-action norm represents, there may be separate constraints on which sorts of state-given reasons for intention are legitimate. Pink argues that there is a constraint on reasons for intention that stems from the idea of intention as having a formal aim to direct the agent towards performing justified actions[60]. This constraint is called *Reason-Apply*:

> *Reason-Apply:* any end E that justifies deciding to do A must, supposing that decision is taken, also provide at least as much justification for doing A.[61]

This is the idea that an intention is justified only if the action at which it aims is also justified by whatever it is that justified the intention to act: that the reasons for intention must have correlates in reasons to perform the corresponding actions. One of Pink's arguments for this constraint is that only if this is true will it be guaranteed that if an agent rationally forms an intention to act, they can then act rationally from that intention:

> Rationally taken decisions to act must leave agents disposed to act rationally…Now for that to be true, the rationality of deciding to do A must guarantee the existence of ends sufficient to justify doing A thereafter… Clearly, it cannot be the end of executing a rationally taken decision… Justification for doing A, then, can only come from what made the decision rational in the first place - from whatever ends provided sufficient justification for deciding to do A. Once the decision has been taken, these same decision-justifying ends must also provide at least as much justification for doing A.[62]

I shall suggest in the next chapter that there is something fundamentally plausible in the idea that rational decisions require the possibility of acting rationally from them. For an agent who knew that it would be irrational to act on some decision can rely on themselves not to carry out that decision, provided that they expect themselves to be rational in this respect (and not, for example, weak-willed). An agent in such a position would be deciding to do something that they expected not to do – and there is something absurd

---

[60] The aim of intention is expressed in the following passage: 'the will is also a reason-applying or executive faculty: its function is to apply reason as it concerns our subsequent action - thanks to the motivation-perpetuating influence of the will, rational decisions to act leave us disposed to act rationally thereafter.' Pink 1996, ch. 5, p.137

[61] Pink, ibid. ch.5, p.153

[62] Ibid., p.152

here. This is captured in Pink's suggestion that decisions be 'motivation-perpetuating' – that, once taken, they allow the motivation to do A to persist right up until the doing of A.

How can *Reason-Apply* be connected to ideas we have already discussed of intention's having a formal aim related to the idea of its producing good or rational action? Pink suggests the following:

> Decision-making, if means-end justifiable at all, must be a *special-purpose* action. That is, not every desirable end which a given decision to act would further can provide any justification for taking that decision. Justification for the decision can only come from those ends which satisfy *Reason-Apply* — which would justify acting as decided thereafter. And the ends which can motivate taking a decision must correspondingly be limited. Any possible motive for taking a decision must *ipso facto* be a motive for acting as decided thereafter.[63]

A further characterization is provided later:

> Deciding to do A, I claim, is an activity - a second-order action - which, whenever it occurs, must be motivated by an overall desire for a specific end. The end in question is simply that the agent does A. A decision to do A implies an exercise of the will which is motivated by a desire that one subsequently does A.[64]

This second characterization explains why the focus in the proof of the partial admissibility of state-given reasons for intention is on the efficacy of a decision in ensuring subsequent performance of the action decided upon. This is because, if an agent cannot achieve A even if they decide to A, the justification for the decision vanishes on this conception. In fact, factors bearing on the efficacy of decision are the *only* kind of state-given reason permitted in this framework: they express the only way in which the desirability of doing A can bear on the wisdom of deciding to do A in a distinctively state-given way. This second characterization also relates much more clearly and directly to the ideas examined in the last chapter to the effect that intention aims at good action: on this conception, intentions are justified only in relation to the desirability of the thing decided upon.

Suppose that a rational action is, at least, an action in whose favour justification of some sort can be given: then *Reason-Apply* entails that any

---

[63] Ibid., p.156
[64] Ibid., p.252

rational intention is an intention on which the agent can rationally follow through – for a rational intention is made rational by the fact that it has substantial justification, and if that very justification must be one which also justifies the intended action, then the intended action is also justified in that way. If this is a correct interpretation of *Reason-Apply*, then the rationale for the 'at least as much' condition within it is a little unclear – as long as the action ensuing from the decision has *some* or enough justification, this should be enough to make it rational[65]. It would not be necessary to give it just the same amount of justification as the decision in order to ensure that its performance could be rationally motivated. This is a key point, and I shall press it in what follows.

The idea of special-purpose agency, however, deserves some scrutiny of its own, since it is supposed to be the underpinning of a conception of decision that excludes certain state-given reasons for decision and intention. The concept of decision as an action that can only be performed for particular reasons itself does not, of course, imply any more particular restriction on which reasons are eligible to govern decision. Even those who support the total exclusion of state-given reasons for intention could endorse a conception of decision as a special-purpose action: they could suggest that a decision to φ could only be taken on the basis of the reasons in favour of φ-ing.

The question is whether the idea of special-purpose action illuminates the nature of the restriction on reasons for decision specified by *Reason-Apply*. After all, if decision is an action, then it is natural to think of it as an action that is up to the agent to perform – and constraints such as *Reason-Apply* become puzzling (thus, if there were no constraints, an agent may have whatever intentions they desire by deciding as they wish). The idea that

[65] It is natural to interpret this 'just as much' condition as attempting to provide for the idea that a justified intention to act is an intention to perform an action that is either the best action or that is at least acceptably good relative to the agent's other possible actions. For its force is to exclude cases in which, though the justification for the intention is there, the justification for the action is also there, though diminished. But once we are comfortable with the idea that a rational action is sometimes one for which the agent has less justification compared to the other actions available to them (as suggested below), it is unclear why we would then exclude such cases as practically irrational: again, as long as there is some justification available for the intended action, then it is unclear why we would not count it under the heading of actions that are rationally performed from an intention rationally adopted.

decisions can only be made for the sake of producing the actions decided upon is the idea of actions that can only be rationally performed for certain reasons. Yet this is not because agents are *incapable* of making decisions for other reasons, but rather because they would not be rational if they did so – because considerations bearing upon whether to produce the action decided upon are the only legitimate reasons justifying decisions.

However, this raises the question why, if the agent's formation of intention is really up to them, they should not be able to employ it when doing so would be useful to them. How can we make sense of the idea that an agent's decisions are actions that are up to them to take, and that that agent has established that taking a particular decision would be useful to them all things considered, but still, they would be criticisable, or confused, if they took that decision?

Pink stresses that the idea of special-purpose action (in application to the theory of decision) ought not be *sui generis*:

> We do not want *Reason-Apply* to be an arbitrary constraint on decision justifications without any parallel in the rest of practical reason. Our theory of decision rationality must not be *ad hoc.* It must not rely on principles of practical reason which are mysteriously specific to the will.[66]

This would be a weakness of the conception of decision as special-purpose agency: it would throw into doubt whether special-purpose agency as such exists (and *a fortiori*, whether decision could be an example). Yet, I will argue, Pink's supposed parallels in the rest of practical reason do not particularly help in this regard. The suggested illustrative examples are from putative other instances of 'end-specific' action: action that 'must, when performed, be motivated by the desire to attain a specific end'[67], just like decision is supposed to be motivated by the desire to do what it is a decision to do. It is correct that if end-specific action exists, special-purpose agency exists, but I will now argue that the end-specific action at play in Pink's examples is not of the same kind as that which is supposed to be involved in decision-making.

---

[66] Ibid., p.248
[67] Ibid., p.248

Pink's main example of an end-specific action is: playing loud music in order to annoy the neighbours. Now it seems that someone could be paid to annoy the neighbours (perhaps by an old enemy of theirs), and to achieve this through playing loud music. If that's right, then they would then count as playing loud music to annoy the neighbours (to get the money), and then the action is not special-purpose in a sense that would exclude external reasons of any sort to take this sort of action, since it is indeed something you can do for money. The point, however, is that a constraint on the end behind the action does occur at the first level: you don't count as playing loud music in order to annoy the neighbours unless your desire is to annoy the neighbours, no matter what in turn justifies your annoying the neighbours. The end-specific action Pink has in mind, then, is playing-loud-music-to-annoy-the-neighbours, which implies aiming at annoying the neighbours. This action is end-specific in one way: specific to the end of annoying the neighbours.

However, this points to a dissimilarity between this sort of case and the treatment of decision suggested by *Reason-Apply*. For this music-playing action is special-purpose for purely *conceptual* reasons: we simply would not describe an agent as playing loud music in order to annoy the neighbours unless they were motivated to annoy the neighbours. In contrast, as we have already noted, the sense in which Pink supposes *decisions* to be special-purpose is that, if decisions are not taken for reasons conforming to *Reason-Apply*, the agent is not making a decision *rationally*. An agent who decides to φ because they will be given money for deciding to φ, but no money for φ-ing, is an agent who (according to Pink) does not possess true justification for their decisions[68].

This difference is decisive. If decision were similar to the case of playing loud music to annoy the neighbours, then the diagnosis would have to be that such an agent who took money for a decision to φ irrespective of the merits of φ-ing would not even count as making a decision to φ – in the same way that an agent who purportedly took money to play-loud-music-to-annoy-the-neighbours, but who is not paid to annoy the neighbours, just would not count

[68] Cf. ibid, p.147-9

as really playing loud music for the sake of annoying the neighbours. But decision is not like this: the barrier here is not conceptual. So this sort of special-purpose action does not substantiate the claim that the *Reason-Apply* restriction on legitimate reasons to decide has parallels in the rest of practical reason. It is not a parallel of the right kind.

So in order to see Pink's conception of decision-making as a special-purpose action as something other than *ad hoc*, we must look for better parallels: we must see whether there are any ordinary examples of actions that are special-purpose in the sense that, simply in virtue of the kind of action they are, they can only rationally be performed for certain reasons. Some actions seem at least to approximate to this category. For example, pursuing a relationship with someone is something that is perhaps only reasonably or justifiably done for certain reasons, where that partially reflects the kinds of things relationships are. Pursuing a relationship with someone solely in order to tick a box on the life-goals list, and irrespective of the value of that relationship, is perhaps very difficult to defend. But even our judgments here will reflect a wide range of implicit psychological understanding and welfare-related concerns: much more than the understanding of what pursuing a relationship simply *is*.

Special-purpose actions that are special-purpose in the rational, rather than conceptual, sense don't appear to exist (or at least, it is very difficult to think any up). If this appearance is correct, then if decisions are such actions, they are *sui generis*. This is not a decisive objection to Pink's conception of decisions, or to *Reason-Apply*, but the difficulty of finding parallel examples of rationally special-purpose actions reflects the puzzle suggested earlier. If decisions are actions, then it makes sense that they are up to the agent to freely take or not take, and though there may of course be restrictions on which decisions may be taken rationally, it is difficult to see why certain reasons would be, as such, excluded, especially if responding to them would be of benefit to the agent. The agent who decides to φ purely on the basis of the monetary award offered for deciding is just not obviously irrational.

While most of the objections to the rational admissibility of state-given reasons for intention to practical deliberation are blanket objections, Pink's *Reason-Apply* presents a unique intermediate case, where only some state-given reasons for intention are accepted. I have suggested so far that the idea of rationally special-purpose agency does not have a parallel in ordinary practical reason. The next section attempts to counterexample *Reason-Apply* by devising acceptably ordinary cases in which justification for a decision (in the positive sense) exceeds justification for the corresponding action.

Beyond *Reason-Apply*: some relevant cases

The essence of *Reason-Apply* is that any justified intention is an intention to perform a justified action – or more precisely, that whatever justifies an intention must justify its associated action, justify it 'just as much'. It is refuted if there are cases in which an intention is justified by more weighty reason than is the action it is an intention to perform: where the action does not make the same kind of contribution to the end as the intention does. In this section I describe two cases.

It may seem easy in general to construct such cases following the formula used earlier: whatever the special psychological effects of intention, there must be cases in which those effects are particularly worthwhile; the fact that the intention has those effects then generates state-given reason to adopt that intention. However, this isn't enough to convince those who deny the rational admissibility of such reasons to practical deliberation, because they may insist that that shows only that an intention with psychologically useful effects is a good intention – not that that intention may rationally be adopted on that basis. (In effect, they would treat such cases as on a par with belief in the existence of God within Pascal's wager). For any case to work it is essential that it strongly seems rational to simply adopt the relevant intention, and not rational merely to take whatever indirect measures are available to ensure that

one ends up adopting the valuable intention – such as by taking a 'decision drug'[69] whose effect is to get oneself to take some specific decision.

Of course, decision drugs don't exist, and neither does any general-purpose measure for getting oneself to adopt any arbitrary intention. That means that a case that seems realistic and mundane enough, where some intention is more valuable than another for state-given reasons, ought to be argumentatively effective. This is because, if we have the intuition that one intention is the rational one to adopt, and that intention is justified by its state-given properties, this intuition then indicates that we would expect a rational agent in such a case to adopt that intention directly, not with the aid of decision drugs or indirect measures.

One kind of case that would outright refute *Reason-Apply* is a case where some intention to φ (where φ-ing is in the future) is rewarded where φ-ing itself is indifferent or bad. The case where φ-ing is undesirable and an intention to φ later is rewarded is just that represented by the Toxin Puzzle (discussed in the next chapter) – a case in which, as we shall see, many philosophers have rejected the rationality of forming the intention whose formation is rewarded. However, the case where φ-ing is itself indifferent is not often discussed, and it would take the following form: the agent is offered a large amount of money to intend to do something at midday tomorrow that is indifferent in its value – say, to drink a glass of water[70]. Whereas forming the intention to drink the glass of water contributes to the agent's financial ends here, actually drinking the water at midday tomorrow does not contribute in the same way to the agent's finances. Now insofar as the reasoning (examined in the next chapter) against the rationality of forming the intention to drink the toxin in the original Toxin Puzzle depends on the inadvisability of drinking the toxin once midday comes around, this simply does not apply to the glass of water case, since drinking the glass of water is not positively inadvisable – it simply makes no difference either way. It is a wholly neutral action. It is not obviously irrational to decide to drink a glass of water at

[69] Ibid., p.96
[70] Cf. Hieronymi 2006, fn. 38 for a brief discussion of this case.

midday tomorrow – after all, why not do that? – especially if doing so would bring a large financial reward.

In fact, given the 'just as much' condition in *Reason-Apply*, there is an even more generous version of the case which is inconsistent with *Reason-Apply*: a case where the agent is rewarded for forming an intention to do an action that is positively useful, but where the action itself contributes less to the agent's ends than the formation of that intention itself.

Here is one such case: the *Incentivized Child Case*. Suppose some parents are attempting to train their child to spend their time well, so they adopt the following practice: every day, they reward the child with something the child greatly desires (such as time with the games console) just in case, at the beginning of the day, the child has formed a specific intention to spend most of their time that day in a way the parents approve of – through specific plans for studying, the performance of specific household chores, or whatever else. The child believes that these things are genuinely good for them to do, but does not desire to do them as much as they desire to spend time playing video games. So they form an intention every day to spend lots of time studying, in appropriately specific ways, that day; their parents approve of each of these intentions and once they believe in the morning that some such intention is genuinely there, they loosen the parental restrictions on the console before going to work to allow their child the allotted time on it. It is very plausibly rational for the child to form a specific intention to do the useful things their parents approve of in order to obtain the gaming time that they more intensely value – the child is a terrible liar, and if they only pretended to have that intention to study, their parents would be able to tell. And what it is an intention to do is something the child would have some reason to do anyway – in fact, once they have played through their allotted gaming time, studying will then be, in their eyes, a desirable enough thing to do, if not a very attractive one.

Here, the studying does not contribute to the child's goals as much as forming the intention to study does. Though their ends justify studying, they do not justify it to the same degree as they justify forming the intention to study. This

makes this case a clear exception to *Reason-Apply*, and it is surely mundane enough to avoid the criticism that the child ought merely to employ indirect measures to get themselves to have the intention to study. They ought, rather, simply to form an intention to study in appropriately specific ways; that is what is intuitive. They ought to form this intention in order to get their desired time with the games console – perhaps in addition to being well-advised to form a specific intention to study because studying itself is worthwhile. That is the ultimate rationale for the parents' adopting this practice: that the child's motivation to play video games will induce them unfailingly to make plans for studying, in a way that their recognition of the worthwhileness of studying wouldn't.

This case counterexamples, in particular, the 'at least as much justification' condition in *Reason-Apply* – the idea that justified intentions are no more justified than the corresponding actions they are intentions to do, and correspondingly that agents are just as motivated to perform the action intended as they are to form the intention (this must be true for the additional reason that if doing A itself motivates a decision to A, as it does according to Pink, then there must be a parity of motivational strength if the agent is rational).

This case does not refute the basic idea of *Reason-Apply*: that rationally taken decisions leave the agent able to act rationally, or at least not irrationally. But it does target the conception of decision as eligible to be rationally taken only for reasons that relate ultimately to the worthwhileness of its object, even if through indirect state-given ways, such as the efficacy of the decision-initiated intention in achieving that object. It challenges Pink's conception of decision as an action that purely applies reason in relation to the actions decided upon, even if through partly state-given ways.

Here is another case that is more similar in its essential structure to Pink's case of Dan the stuntman – I will refer to it as the *Train Case*[71]. A group of

[71] A simpler version of the case can also be mounted against the blanket denial of the admissibility of state-given reasons: some holidaymakers need to leave at 7.30, but suspect they may be ready to leave too late, so they decide instead to leave at 7, thinking that they can thereby assure themselves of leaving at the latest by 7.30.

holidaymakers have come to the end of their holiday and would prefer a relaxed and pleasant journey home. They are obliged to catch the ferry home (only one ferry leaves per day). The ferry leaves at 10 o'clock, and to get there they must take a train to the port. The relevant train (there is only one in the morning) leaves at 8 o'clock, so as long as they leave the house by 7.30, they will get there on time. Moreover there is a fair amount of stuff to do before they go – last-minute cleaning, locking up and so on, as well as final packing and gathering of stuff. So they will have to get up at least an hour before they intend to leave in order to do it all. This means that they will have to get up by 6.30 at the latest.

Now, they are a big and chaotic group, and they are somewhat likely to be ready to leave later than they intend. If they intend to be ready to leave at 7.30, then they face a substantial risk of not being ready by then and so not catching the 8 o'clock train, and this would leave them stranded – a highly undesirable outcome. But if they intend to be ready to leave earlier, then they must accordingly push back the time they are to wake up, and the further they push back the time, the groggier and more irritable they know they will be. Most probably, they will sleep in a little relative to the time they intend to get up. However, they also know that even if they are ready to leave too late, they won't be more than half an hour late on this score. So they are certain that if they plan to get up at 6 o'clock in order to be ready to leave by 7 o'clock, they will leave on time to catch the train to the port. If the holidaymakers are rational, they will surely intend exactly that, since an intention to leave later brings avoidable risk – even though the intention to leave later is justified better by its object-related reasons, since getting up at 6.30 and leaving by 7.30 fulfils better the goal of a pleasant journey home.

This case is another counterexample to *Reason-Apply*, specifically in relation to the 'at least as much justification' condition. Getting up at 6 and being ready to leave by 7, that course of action, fulfils their goal of a pleasant journey home less well than the course of action of getting up at 6.30 and being ready to leave by 7.30 – not that much difference is made, but it is enough so that the latter course is preferable. The former course of action, if actually taken, would leave them with (let's say) three units of grogginess.

But since they are somewhat likely to implement their intentions unpunctually by sleeping in a little, they are likely, if they intend to get up at 6, to get up at 6.30 by the latest. So intending to get up at 6 is something that will probably result in them getting up at a time that will leave them with (let's say) two expected units of grogginess. This means that the intention to get up at 6 o'clock in order to be ready by 7 fulfils their goals better than the course of action thereby intended (actually getting up at 6 to be ready by 7). This makes it a counterexample to *Reason-Apply*, taking the 'at least as much justification' condition at face value and interpreting it in the way I have done.

This case does has the feature pointed out with respect to Dan the stuntman: it is open to us to construe the intention to get up at 6, not as more justified than the course of action thereby intended, but rather as the only acceptable intention once the other possible intention, intending to get up at 6.30, is rejected because it cannot ensure the discharge of the agents' ultimate aim of getting home. If that's right, then it only confuses the description of the deliberative situation to imagine the agent picking a superior intention while recognizing that its justification exceeds the justification of the action decided upon. We have here only a negative state-given reason, a reason against an intention, not a positive state-given reason in favour of one.

Another issue with the case is that it is tempting to try to redescribe the agents' possible intentions as intending to *try* to get up at 6 in order to be ready by 7, and intending to *try* to get up at 6.30 in order to be ready by 7.30. After all, it has been conceded that they are at least somewhat likely to fail – so perhaps the language of trying is appropriate here (this conception of trying is discussed more in chapter 5). Such intentions would mirror in their justification the courses of action intended: trying to get up at 6 is precisely the activity that the holidaymakers ought to pursue, because it guarantees passage home with two expected units of grogginess, which is the same that can be said for the intention. Nonetheless we may suppose that leaving at the time they intend is basically within the group's power, and that their concern is only with *assuring* themselves that they will leave at the latest acceptable time. Given that, there is no justification for describing the agents as only *trying* to leave on time. They expect to leave at the time they choose, but only

want, quite reasonably, to avoid incurring the danger of a terrible worst-case scenario.

This case exploits not the psychological power of decision-making or of the implementation mode that can accompany intentions, but rather a particular psychological and knowable fact that pertains to the agents, namely, a predictable sort of unreliability with respect to whatever it is that one intends. This regular unpredictability with respect to the effects of an agent's intentions is also what Pink's case of Dan the stuntman relies on. It is because Dan is wary of a foreseeable preference shift on his part that he decides to adopt what would otherwise be a second-best plan. That is also true of the holidaymakers: they expect a temporary preference shift at the moment of their alarm going off in favour of sleeping more, and they too adopt what would otherwise be a second-best plan. So this case trades on the same rationale as the case of Dan the stuntman. The only difference is that, whereas Dan's doing both does fulfil his fundamental goal of avoiding doing only one or the other just as much as his *plan* to do both, in the *Train Case* the actual course of action of getting up at 6 in order to leave by 7 fulfils the holidaymakers' goal of a pleasant journey home slightly less well than their plan to do that. This is, in effect, a tweak, but its significance is that it shows the independence of two factors in Pink's case of Dan the stuntman: the state-given reasons that justify the decision to both organize publicity and attempt the stunt, and the fact that doing both happens to contribute to the agent's goals just as much as that decision to do both does. It is because these factors are independent that cases can be devised in which they come apart; and the Train Case is one such case.

Conclusion

If these cases are compelling, then they suggest that state-given reasons for intention are rationally admissible even where they don't relate to any putative formal aim of intention with respect to good or worthwhile action. They present a reasonable initial case for the legitimacy of such reasons. To defend their existence fully, it is necessary to combat various skeptical

challenges – such tasks are taken up in subsequent chapters. These cases are successful if they force the theorist hostile to state-given reasons to recognize that the onus is on them to show why state-given reasons for intention are, despite appearances, rationally inadmissible to practical deliberation.

3. Neutralizing the Toxin Puzzle

So far, I have addressed some objections to the possibility of state-given reasons for intention, and I have provided some cases in which it seems clear that an agent is justified in forming an intention despite not having reason, or reason of corresponding strength, to perform the action that would then be intended. In this chapter I shall address a further argument against the possibility of state-given reasons for intention: that to accept their possibility would commit the theorist to an unacceptable conception of what is justified, or rational, of an agent who finds themselves in a choice situation whose justification structure is characterized by Kavka's Toxin Puzzle[72]. As before, I will continue to use the term 'state-given reasons' to refer to reasons to intend that stem from the desirable properties of the intention that are not related to the usefulness of the action that is the intention's object.

I won't run through the puzzle in its usual form; I shall assume the reader's familiarity. To fix ideas, I shall stick with characterizing the puzzle abstractly. The type of justification structure of actions and intentions specific to the Toxin Puzzle is characterized by three essential points[73].

First, there is some future-direction intention whose formation is highly desirable, where the explanation of this desirability has nothing to do with the intention's likelihood of bringing about the object of that intention. In the original puzzle, the intention is to be formed by midnight; if the agent intends at midnight to drink a mild toxin tomorrow at midday, the agent receives a million dollars.

Second, this intention ceases to be positively desirable at some point in between the hypothetical deliberation conducted by the agent at the centre of the puzzle and the time at which the intended action is to be performed. In the

[72] Presented originally in Kavka 1983.

[73] This structure is not unique to the Toxin Puzzle and also characterizes the traditional paradox of nuclear deterrence (cf. Kavka (1978)). It sheds light on the latter insofar as it reveals that the temporal structure of the deterrent, whereby the firing of the nuclear weapons would have to occur after any benefit from the intention (to fire them in certain circumstances) was exhausted, is the real issue for the plausibility of that intention – not the atrocity inherent in such use itself.

original puzzle, the money is received the next morning depending solely on whether or not the agent intended at midnight to drink the toxin at midday the next day. After midnight, there is no longer anything gained in having that intention. This feature of the Toxin Puzzle, incidentally, marks a contrast with some of the cases used in the previous chapter to prove that intentions can be justified in their own right; those cases were constructed so that the intention remains desirable to have at the time of the action to be performed from it.

Thirdly, the action that the desirable intention is an intention to perform is itself undesirable to actually perform; the agent is better off if they don't perform it. But it is not so undesirable to perform as to outweigh the reasons there are to intend to perform it, to the extent that those two sets of reasons are in competition. In effect, the unique rational preference ordering for the Toxin Puzzle is:

1. Intend to drink the toxin at t_0, don't drink the toxin at t_1

2. Intend to drink the toxin at t_0, drink the toxin at t_1

3. Don't intend to drink the toxin at t_0, don't drink the toxin at t_1

4. Don't intend to drink the toxin at t_0, drink the toxin at t_1

What is contestable is whether this rational preference ordering corresponds to what it would be rational for the agent to do, or intend; as we will see, many philosophers claim that an agent in such a situation could not rationally intend at t_0 to drink the toxin at t_1. This is because of the crux of the puzzle: the point that when midday comes, the agent is better off not drinking, so they cannot intelligibly choose to drink the toxin then, or to consider drinking it a good thing – and so they cannot choose in advance to drink the toxin, which is what intending at midnight to drink it amounts to. If this is correct, this discrepancy with the apparent rational preference ordering is remarkable in its own right – and it stands in a clear tension with our suggestion that the formation of intentions is made rational by the state-given reason to have those intentions, since the financial reward creates, it seems, ample reason to have the intention to drink the toxin. Exploring this tension and assessing its significance is the task of this chapter.

One of the thoughts explored in chapter 1 was the idea that, in practical deliberation, we just do focus on what we shall have reason to do, and we accept a norm that our practical deliberation not only is, but is properly, determined by such considerations alone. This thought has been expressed independently in discussions of the Toxin Puzzle by philosophers whose intuitions are strongly in line with the position that the intention to drink the toxin is an irrational one, despite the massive financial reward for having that intention. For instance, Bratman writes:

> [One relevant assertion is] that your reasons [at midnight] for intending to drink the toxin [at midday] will make it rational of you so to intend on the basis of deliberation. But in deliberation about the future we deliberate about what to *do* then…This means that in such deliberation about the future the desire-belief reasons we are to consider are reasons for various ways we might *act* later. It follows that your special million-dollar reason for intending now to drink the toxin later will not get into the deliberation about whether to drink it later.[74]

It is clear that Bratman thinks there is only one single kind of practical deliberation in the Toxin Puzzle: deliberation about what to do in the future. If the answer to that deliberation is 'drink the toxin', the agent then intends to drink the toxin, whereas if the answer is 'don't drink', then the agent shall not so intend. From this perspective, the rewards for having that intention now are irrelevant to the central question of practical deliberation so defined.

This line of argument seems to embody a confusion between two kinds of deliberation: deliberation about what to intend now, and deliberation about what to do later. Even if deliberations of the latter sort would lead the agent to form certain intentions that are possibly inconsistent with intentions they have gained from the former kind of deliberation, it still does not follow that the two deliberations are one and the same. In conflating these two kinds of deliberation, Bratman has arguably begged the question when we says that we solely deliberate about what to *do* in the future (rather than what now to intend to do later).

To the extent that all the Toxin Puzzle does is to make the intuitiveness of Bratman's line of thought vivid, then there is no special need for us to discuss it in the context of defending the existence of state-given reasons for

[74] Bratman 1987, ch. 6.6, 'Kavka's Puzzle', p.103.

intention. All that is needed to dispel this intuition is to refer the reader to a wider variety of cases, including the ones discussed in the last chapter. Those cases strongly suggest that deliberation on what now to intend must be logically separable from deliberation solely on what shall be the most preferable thing to do later, and moreover, that the latter kind of deliberation does not always fully determine the proper results of the former. How it is rational for agents to behave in those cases cannot be accounted for unless we suppose that agents may deliberate on what to intend in its own right.

However, dealing with Bratman's response to the Toxin Puzzle requires more than citing this confusion. Two things need to be said. The first is that it is itself important that Bratman is able to cite this intuition; such intuitions ought not be simply dismissed. Something must explain why various writers are drawn, in the context of the Toxin Puzzle, to the idea of the primacy of action within the context of practical deliberation. Given that other cases of the kind adduced in earlier chapters illustrate the opposite idea, it seems that if it is true that practical deliberation *can* be concerned with what to intend in its own right, then there ought to be no obstacle, in general, to concerning oneself in the context of practical deliberation with what to intend. If that's true, then why is it not intuitive in the case of the Toxin Puzzle that the agent ought at midnight to concern themselves with whether to intend to drink the toxin?

It would be theoretically unsatisfying to write off this intuition as mere dogma. It would be better if we could explain something of the appeal of the intuition, even within the broader context of arguing that it is ultimately misguided. Whether or not an explanation of that kind is available can be used to judge the view. We will return to this issue at the end of this chapter.

The second point is that there is a more sophisticated argument that Bratman makes that does not simply conflate the two kinds of deliberations. We shall examine this argument later in more detail, but for now it suffices to note the two basic premises (which are deployed in this context in (1998))[75]. The first

[75] Bratman, 'Following Through with One's Plans: Reply to David Gauthier', in Danielson (ed.) 1998.

is that it is not rational for the agent actually to drink the toxin at t_1; it is rational for them to refrain. The second is the 'linking principle':

> I can rationally decide now, on the basis of deliberation, on a plan that calls for my *A*-ing in certain later circumstances in which I retain rational control over my action, only if I do not now believe that when the time and circumstances for *A* arrive I will, if rational, reconsider and abandon the intention to *A* in favour of an intention to perform some alternative to *A*.[76]

From these two principles it follows that it is not rational for the agent to intend at t_0 to drink the toxin at t_1. And this argument in no way conflates two distinct kinds of deliberation; rather, if anything, it respects that distinction, by suggesting how deliberation on what to do later can rationally constrain deliberation on what to intend now. And this leads to the clear tension with the rational admissibility of state-given reasons for intention: for given the rational preference ordering suggested earlier, it seems that if state-given reasons make a difference at all in matters of rational intention, then they should make rational the intention formed at midnight to drink the toxin at midday.

In the face of this conclusion, one option for the defender of the rational admissibility of state-given reasons for intention is to try to defend the rationality of drinking the toxin at midday in the circumstances of the Toxin Puzzle and of an agent's appropriate response to it. This position has been propounded in particular by David Gauthier[77]. There is a clear rationale: if, given the situation at midnight, it were rational to drink the toxin at midday, it would be rational to form at midnight the intention to drink the toxin. If it were so rational, then a rational agent could win the money. Since winning the money is *ex hypothesi* preferable to merely not drinking the toxin, such an agent would be in the best position to fulfil their rational preferences, which gives them a clear claim to rationality.

Gauthier's position has been ably criticized[78]. The underlying issue is that the implicit picture of rationality Gauthier is using itself is the one used to show that drinking the toxin is itself (*contra* Gauthier) irrational. For Gauthier's

[76] Ibid., p.55.
[77] Cf. in particular Gauthier 1984; Gauthier 1998 (in Danielson, ibid.) for a revised view.
[78] Cf. Levy 2009; Pink 1996. ch. 6

position suggests that it is in virtue of the fact that intending to drink the toxin is a dominant strategy relative to not so intending that it is rational to so intend. This reasoning – that when, of two alternatives, one is superior, then performing it is the rational course – is exactly that used to justify not drinking the toxin. For Gauthier's strategy to succeed, it must show that this principle is somehow neutralized in respect of the act of drinking the toxin, but not so neutralized in respect of which intention to adopt. Yet this principle seems to be fundamental to practical thinking. It is difficult not to see modifications to this principle as *ad hoc* – as amounting ultimately to the idea that, if the agent conducted practical thought in a different way, they would be better off.

A full examination of Gauthier's views is beyond the scope of this chapter. This chapter will not defend Gauthier and will not attempt to suggest that actually drinking the toxin, even in the circumstances of the Toxin Puzzle, is rational if the agent forms at midnight the intention to drink it in order to win the money. Instead I shall focus on a different aspect of the problem here for state-given reasons for intention. The relevant argument is the following. I shall call it the *Anti-Desirability Argument*, because of its conclusion: that the desirability of an intention does not create reason to adopt it. On this argument, if the desirability of an intention creates reason to adopt it, then an agent, in a Toxin Puzzle-like situation, has reason to form the intention to drink, and so it is rational for them to form that intention. Yet it is not rational for them to have that intention; therefore, they cannot have reason to form the intention to drink. In turn, this shows that the desirability of that intention, on account of its own properties as a state, does not suffice for the existence of reason to form the intention to drink. The desirability of an intention, in this case, does not create reason to adopt it. So, generalizing, there are no state-given reasons for intention (or at least, 'reasons' in the sense of 'reasons that are rationally admissible within practical deliberation' don't exist, as opposed to factors making it good to have certain intentions).

Writers on the Toxin Puzzle have often been concerned only with whether it is rational to intend at midnight to drink the toxin; the further issue, of state-given reasons for intention, has not been discussed as much. The Anti-Desirability Argument spells out this connection. This argument has been

made by Pink (in the context of accepting the principle that rational intentions leave the agent able to act rationally from them)[79]:

> not every desirable end which a given decision to act would further can provide any justification for taking that decision. Justification for the decision can only come from those ends…which would justify acting as decided thereafter. And the ends which can motivate taking a decision must correspondingly be limited. Any possible motive for taking a decision must *ipso facto* be a motive for acting as decided thereafter…So prizes offered simply for deciding to do A, no matter how large they might be, cannot provide any justification whatsoever for taking that decision, or motivate us to take it. And that is because those prizes cannot do anything to justify or motivate doing A thereafter.[80]

This allows us to connect the Toxin Puzzle to a puzzle raised in the previous chapter: if intentions are genuinely up to the agent to form, then agents should be able to form them for whatever reason they please. Kavka makes this point too: 'You are asked to form a simple intention to perform an act that is well within your power. This is the kind of thing we all do many times a day. You are provided with an overwhelming incentive for doing so.'[81] In this light, the desirability of the intention to drink the toxin plainly ought to make intending it intelligible. The fact that it doesn't would then suggest that intentions, and decisions, aren't really up to us in the usual way.

The fundamental difficulty in the Toxin Puzzle is that two apparently plausible ideas run into conflict. The first is that the agent should be able to win the money by forming the intention. After all, forming the intention is a dominant strategy and, since its formation only reflects the agent's own agency in this matter, it seems that it should be well within the agent's power to win the money by forming the intention. The second plausible idea is that the agent must be able rationally to refuse to drink the toxin when the appointed time comes – indeed that they are rationally required to not drink the toxin. Since drinking the toxin in no way benefits the agent relative to not drinking the toxin, and the agent knows all the facts that show exactly that, it

[79] Ibid., p.152
[80] Ibid., chapter 5 'Decision Rationality and Action Rationality', p.156
[81] Kavka 1983, p.35

seems that the agent must conclude that they ought not to drink, and act on that judgment.

One possibility is that the conflict between these two ideas may be irresolvable. Kavka writes:

> When we have good reasons to intend but not to act, conflicting standards of evaluation come into play and something has to give way: either rational action, rational intention, or aspects of the agent's own rationality (e.g. his correct belief that drinking the toxin is not necessary for winning the million.[82]

The suggestion that, in Toxin-Puzzle-like circumstances, agents are incapable in principle of achieving full rationality, or meeting reasonable standards of justification, is surely an unattractive one. It makes more urgent discussion of the Anti-Desirability Argument, which, along with the principle that rational intentions leave the agent able to act rationally from them, is the crucial connection in establishing the tension between state-given reasons for intention and the irrationality of drinking the toxin.

I shall divide my discussion of the Anti-Desirability Argument into discussion of its main premises:

1) It is not rational in Toxin Puzzle-like situations to form the desirable intention (in the Toxin Puzzle itself, the intention to drink the toxin).

2) If it is not rational to form an intention, an agent lacks overall reason to form that intention.

3) If in Toxin Puzzle-like situations an agent lacks overall reason to form the intention to drink, then they lack any reason to form the intention to drink.

4) If in Toxin Puzzle-like situations an agent lacks a reason to form the desirable intention, then *in general* the desirability of intentions does not create reasons (of any strength) to form those intentions.

Conclusion) The desirability of intentions does not create reasons to form those intentions; there are no state-given reasons for intention that stem from intention desirability.

---

[82] Ibid., p.36

Premises 2) and 3) must be discussed first, since the objections they appear to invite can be dealt with quickly, whereas the other two premises require more significant discussion.

Premise 2) asserts that, were the agent to have state-given reason to form the intention to drink the toxin, it would then be rational for them to do so. This may be thought doubtful, however, because a certain generalization of this idea fails. After all, we have already conceded that, for beliefs, the desirability of having the belief would not make a relevant difference in making it rational to have that belief.

To take an extreme example, suppose an agent is offered a large sum of money for believing a contradiction – that Fermat's Last Theorem is both true and false. That they are offered that money manifestly fails to make it rational for them to accept that pair of beliefs. The same point can be made with less extreme examples. Suppose an agent is offered a large sum of money to believe something that they know that they have insufficient evidence to accept – say, to believe that the extant proof(s) of Fermat's Last Theorem contain hitherto unnoticed mistakes and that the theorem is actually false. Again, it seems clear that the financial reward does not make it rational for them to accept this suggestion.

Of course, it is not clear just in virtue of the description of the case whether we should characterize such examples in terms of a discrepancy between what the agent has reason to believe/intend and what it is rational for them to believe/intend, as opposed to a discrepancy between what it is desirable for them to believe/intend and what they have reason to believe/intend. To deny premise 2), in the context of the Toxin Puzzle, would be to use the former as the basis of one's diagnosis.

A denial of premise 2), as a means of blocking the Anti-Desirability Argument and so protecting the existence of state-given reasons to form intentions, suffers from two decisive flaws. Firstly, even if it were a good response to the Anti-Desirability Argument, the position one would be left with would be highly unsatisfying in terms of a theory of the rationality of intentions. For a denial of premise 2) amounts to the assertion that even if

there are state-given reasons to have intentions, and even if those create overall reason to have some intention, it still does not follow that it is rational to form that intention. Yet the cases examined in the previous chapter relied on exactly that link: it is because the incentivized child has good state-given reason to intend to study, and because the holidaymakers have good reason to intend to get up at 6, that it was rational for them to have those intentions. Without a connection such as that expressed in premise 2), this reasoning cannot be carried through, and this creates an explanatory gap in our analysis of those cases. Even if denying premise 2) could block the Anti-Desirability Argument, it would only create problems elsewhere in the system.

Secondly, it is not meritorious in its own right. The idea that rationality on occasion diverges from an agent's reasons derives plausibility from cases of non-irrational ignorance. It reflects intuitions such as the following: that if an agent rationally believes that gin is in the bottle, whereas it is in fact petrol, and they have reason which they are engaged with to get themselves a gin and tonic, then it is rational for them to mix a drink with what is in the bottle, though they do not have reason to use what is in the bottle (they may of course rationally think they have such reason)[83]. The resulting position that what it is rational to intend *can* diverge from what an agent has reason to intend is not universally accepted, but those positions on which it does are positions on which it diverges specifically for cases where some sort of ignorance is in play[84].

If either of these sorts of position are correct, then premise 2) cannot be denied in the context of the Toxin Puzzle. If rationality does not ever diverge from reasons, then premise 2) is correct in application to any choice situation,

[83] The example is of course drawn from Williams, Internal and External Reasons, in Williams 1981, p.102.

[84] The key idea here is that a person can rationally fail to know what their reasons require of them – for an argument here see Broome 2013, ch.5. Some of the relevant issues are reflected in the literature on subjective and objective 'ought': the position that what it is rational for an agent to do and what they have reason to do can diverge can be closely compared to the position that, in some situations, two valid uses of 'ought' in relation to the agent's situation can be distinguished, one that reflects what the agent knows, and the other which simply reflects the actual facts and values; see here Gibbard 2005 for a defence of this view. Kiesewetter 2017 is a recent example and provides a well-developed defence of the position that what it is rational to do is (always and necessarily) what the agent has reason to do. Other strategies are adopted by Jackson and Pargetter 1986, and Kolodny and Macfarlane 2010.

including the Toxin Puzzle, and there is no means here to block the Anti-Desirability Argument. If, on the other hand, rationality can diverge from reasons in cases (or some cases) of rational ignorance, then whether premise 2) is correct in the case of the Toxin Puzzle depends on whether the agent, deliberating on whether to intend at midnight, is ignorant of any relevant fact. But no such thing holds: the agent in the Toxin Puzzle knows exactly the rewards and costs of both intending to drink the toxin and of drinking the toxin itself. They are not ignorant of any particular; there is no room here to get going an explanation of how it could be irrational for them to intend to drink the toxin, even though they have overall reason to so intend.

Premise 3) also requires justification because its general analogue is spectacularly implausible. Lacking overall reason to do something does not in general imply that one lacks *any* reason to do it. The falsity of this thesis is one of the things that makes useful talk of normative reasons at all.[85]

However premise 3) is true of the Toxin Puzzle. The reason is this: if the desirability of the intention to drink genuinely creates reason to form that intention, then, *ex hypothesi*, it creates very strong reason to form that intention to drink. As Kavka puts it: 'you have every reason (or at least a million reasons) to intend to drink'[86]

The reason to intend to drink, if genuine, is stronger than the reason there is not to intend to drink. This is true even if intending to drink will bring about actually drinking the toxin. As per the rational preference ordering described at the beginning of this chapter, it is better to intend to drink the toxin and actually drink the toxin than not to intend the toxin and not drink it. This implies that if the desirability of the intention to drink creates reason to intend to drink, then it creates reason to drink that is stronger than whatever reason there is not to intend to drink. Thus, it implies that the agent has overall reason to intend to drink. Conversely, it is extremely plausible that if the agent lacks

[85] Cf. the Chisholm/Raz exchange in Körner (ed.), 1974. Also related to this issue are old discussions of regret and its apparent capacity to display normative structure: particularly noteworthy pieces here are chapter 2 of W.D.Ross 1930, and Williams, 'Ethical Consistency', in Williams 1983.
[86] Kavka 1983, p.35

overall reason to intend to drink the toxin, then they lack any reason to intend to drink.

To deny this would be to assert that even though the agent has *some* reason to form the intention to drink the toxin, still, some feature of the situation makes it the case that this reason is necessarily less than the reason they have not to intend to drink. But this thesis would imply that the preference ordering above is not in line with what the agent actually has reason to do; but since that preference ordering is clearly justified, this thesis cannot be correct (to my knowledge, no-one argues this).

As mentioned, this chapter shall not attempt to argue that it is in fact rational to intend to drink the toxin (if it is, then the Anti-Desirability Argument fails, and there is just no problem here for the rational admissibility of state-given reasons for intention). I shall assume that premise 1) is correct. This leaves premise 4). In this context, the Anti-Desirability Argument can be blocked only if premise 4) is denied: that is, if a way is found to suggest that, although the desirability of intentions generally create reasons to adopt them, this is not so in the case of the Toxin Puzzle.

The only prospect for denying premise 4) in a principled way, that I can see, is to suggest that it is the fact that forming the intention to drink would be practically irrational that makes it the case that the desirability of the intention does not create reason to form it.

The barebones of such a view can be constructed from the materials supplied so far. The co-incidence of rational requirement and decisive reason, that holds at least in the case of the Toxin Puzzle, leaves open the explanatory direction – whether it is irrational because it is contrary to the agent's reasons, or if is unjustified because it is irrational. Perhaps, in general, there is no conceptual priority of one over the other. What is needed is for it to be the case that on this occasion that irrationality of the intention to drink the toxin explains why that intention cannot be supported by the agent's reasons. The intuitive idea is simply that because the intention doesn't make sense, or is clearly criticisable or absurd, it cannot possibly be justified – so its desirability does not create justification for it.

This is, perhaps, an unusual attitude towards the relationship between reasons and rationality: that the putative violation of a rational requirement explains the absence of justifying reasons. It is opposed to the widespread view that rationality is to be explained in terms of reasons. On Kiesewetter's recent version, this is the view that rationality '*consists in* responding correctly to reasons…that requirements of rationality just are requirements of reasons'[87] (my italics). It is also inconsistent with the more indirect capacity claim, on which 'an account of rationality is an account of the capacity to perceive reasons and to conform to them'[88]. Such views require the conceptual priority of reasons over rationality, and this priority is inconsistent with the way of denying premise 4) suggested here.

Despite these looming objections, I suggest that denying premise 4 in this way is the best option for those who wish both to grant the irrationality of forming the intention to drink the toxin, but who wish also not to deny the general possibility of state-given reasons for intention. But to flesh out what this view might look like, we must explore what sort of independent rational requirement might be breached in the case of the Toxin Puzzle: a requirement that can be understood prior to considering whether the desirability of forming the intention to drink creates a reason to form that intention. We must assess premise 1) in more detail.

The point of doing this is to attempt to discern what rational requirement on intention could explain why the desirability of the intention to drink the toxin should fail to create reason to have that intention. Because the point is to explain why there is no reason to have that intention, our understanding of that rational requirement cannot simply presuppose an account of what reasons there are to have intentions. If it did, then that account would either allow the state-given reasons of the Toxin Puzzle to exist – in which case premise 4) is left standing – or else it wouldn't allow them to exist, in which case it runs the risk of being an account of the kind already rejected, such as one that excludes the existence of any state-given reasons, or that excludes too many state-given reasons. We are looking for a rational requirement on

[87] Kiesewetter 2017, ch.7
[88] Raz, 'Explaining Normativity: On Rationality and the Justification of Reason', in 1999.

intention that has a basis in something genuinely independent of conceptions of reasons for intention. Only that will help us deny premise 4).

The irrationality of intending to drink the toxin: Bratman's argument

One of Bratman's arguments for the irrationality of forming the intention to drink has already been considered and rejected: the appeal to phenomenology. Towards the end of this chapter I will supply the missing explanation of why it should be that Bratman's intuitions are more natural in the case of the Toxin Puzzle than elsewhere. But first I shall consider the other argument made by Bratman that suggests that it is irrational to drink the toxin, an argument that does not rely on the premise that intentions just are determined, if the agent is rational, by what the agent takes themselves to have reason to do.

Bratman's argument is complex, but the core consideration occurs in the following passage (on his version, Tuesday is the equivalent of Kavka's midnight the day before, and Wednesday is the equivalent of the midday the day after when the toxin is to be drunk):

> Given the strength of your reasons for intending on Tuesday to drink the toxin on Wednesday, it may well be rational of you deliberatively so to intend on Tuesday. But your intention to drink the toxin on Wednesday is an intention to drink the toxin under certain expected conditions. Come Wednesday, these are exactly the conditions that you know to obtain. There is no relevant divergence from your expectations on Tuesday about what will happen on Wednesday. But if there is no such divergence from your relevant expectations, it will be rational of you not to reconsider the intention on Wednesday. So on [Bratman's] theory it may well be rational of you on Wednesday intentionally to drink the toxin…And that seems wrong.[89]

Although there are multiple points of entry here for skepticism, Bratman suggests that the best option – at least for his theory – is to deny that it is rational on Tuesday to intend to drink the toxin. The appeal to the phenomenology of practical deliberation discussed earlier then figures as his way to motivate this denial[90]. The denial, strictly, is that it is rational *deliberatively* to intend on Tuesday to drink the toxin: it is irrational, according to Bratman, to adopt an intention to drink the toxin if one is forming

[89] Bratman 1987, 6.6, p.102
[90] Ibid., p.103.

an intention through practical deliberation. That is the result of the assumption that practical deliberation intrinsically is concerned with what to do, not what to intend.

That leaves Bratman with the possibility of nondeliberative intention formed on Tuesday: an intention to drink the toxin formed on Tuesday not through practical deliberation but by some other means (perhaps through some sort of self-deception). In the case of nondeliberative intentions, Bratman suggests that it is false that it is rational of the agent not to reconsider their intention on Wednesday. Although some such principle of rational nonreconsideration is supposed to hold for deliberative intentions, it does not for nondeliberative intentions: 'even if it was rational of you to acquire a new intention by way of such non-deliberation-based processes, and even if circumstances develop as expected, it might still be incumbent on you to reconsider if the opportunity arises.'[91] This assertion then blocks the conclusion that it is rational of the agent to drink the toxin on Wednesday.

The basic consideration has some claim to be a good argument. It certainly seems right that, normally, it doesn't make sense to reconsider an intention when everything is as was expected when we initially formed that future-directed intention. It also seems right that if it's rational to have at time t an intention to φ at t (i.e. then), then it is rational to φ at t. These are the other assumptions made.

if Bratman is right, then this is a rational requirement that is breached by intending to drink the toxin: the requirement not to adopt intentions that it will be both rational not to reconsider and rational not to act on. However, it is not obvious that this rational requirement fulfils our ambition to find a rational requirement on intention which has a basis in something other than a conception of what reasons there are for intentions. The problem is that the rational non-reconsideration premise falls foul of this. I will argue that there are two possible readings of its basis: either it has a basis in a conception of what reasons there are for intentions, or it has a basis in a general prohibition on reconsideration. If the former, then it is not an independent rational

[91] Ibid., p.106.

requirement and so cannot help us deny premise 4); or else it is independent, but then (I will argue) it cannot be given a strong enough basis. It is to be emphasized that this is not an argument *against* the rational non-reconsideration requirement. It is rather an argument against trying to incorporate that requirement into a diagnosis of the deep structure of the Toxin Puzzle and how it does and doesn't create reasons to intend.

Let's first consider Bratman's own understanding of the rational non-reconsideration principle: that the reason why it is normally irrational to reconsider intentions when things are as expected is because it is important in general that plans be stable unless there is some problem for the plan[92]. He clarifies: 'problems for a prior plan will normally involve at least one of the following elements: some relevant divergence between the world as one finds it and the world as one expected it to be when settling on the plan; some relevant change in one's desire or values; or some relevant change in some of one's other intentions.' An agent who reconsiders their plans even when there is no problem with them tends to have unstable plans, and this is contrary to their long-term interests in normal conditions, given the benefits of stable plans[93]. This general fact makes it the case that it is a reasonable habit not to

[92] Ibid., p.66-7

[93] Tenenbaum 2016 spells out why reconsideration is, without some special rationale, unhelpful: 'Of course, it would be a horrible fate to spend most of one's life reconsidering one's decisions; it would be absurd if a theory of rationality required us to use all our leisure time, let alone all our time, in the service of better deliberation. Since the completion of many of our ends would be hampered if we were to deliberate too much, we are all already under a requirement not to reconsider too much. Given that I have among my ends exercising, spending time with the children, watching cartoons, etc., I cannot coherently also engage, at the same time, in endless deliberation…It would also be a sad kind of life spent reconsidering our beliefs all the time. There is thus no rational requirement to reconsider our beliefs at every opportunity, or to always look for further evidence; we are often permitted to settle on a belief.' This justification only applies on balance; it doesn't follow from the potential of reconsideration to interfere with projects that, for any intention, one is by default prohibited from reconsidering that very intention: 'Quite often reconsidering is not too costly, and for (almost) any particular plan, abandoning just this plan will not undermine more general ends. But reconsidering in too many cases can have a devastating cumulative effect: were I to reconsider my intentions at every permitted opportunity, I would forgo pursuing many of the ends I care about. And were reconsideration to lead me to revise plans often enough, my life would be a pathetic alternation of momentary or soon-to-be-abandoned pursuits. On the one hand, no particular intention must "resist reconsideration"; the requirement not to reconsider too much applies only to the total set of one's intentions.'

reconsider one's plan unless there is some problem with it of the sort spelled out in the above.

According to Bratman, agents who exemplify reasonable habits are *normally* rational in doing so. It is occasionally rational to cause oneself to have unreasonable habits, but, according to Bratman, it remains true that the reasonable habits define a standard of rationality for agents[94]. This general conception of practical rationality secures the rationality of nonreconsideration of an agent's intentions in Toxin Puzzle-like situations only by exploiting the fact that Toxin Puzzle situations are not normal. If they were normal, then it would be to agents' long-term interests to be able to negotiate Toxin Puzzle-like situations by forming the rewarded intentions and then reconsidering them later.

We should be wary, however, of identifying the unusualness of the Toxin Puzzle, or its obvious artificiality as a set-up, with its abnormality. Toxin Puzzle-like situations are not that unusual, and they are dubiously beyond the range of normal situations. On Bratman's view, the ultimate standard of 'normality' here, the criterion for how habits of reconsideration are to be assessed, is 'the extent to which these habits come up to a standard appropriate for guiding the education and development of agents like us over the long run.'[95] It follows that as long as Toxin Puzzle-like situations are a contingency that occurs within even a minority of lives, then it may be useful to make the optimal handling of them a standard of development or education. Such standards are surely relatively weak: the question is just whether, though such education, the pupils will be helped. If enough pupils are helped, then it is useful in principle to make the optimal handling of Toxin Puzzle-like contexts part of education. There is no understanding of 'normality' here that could *in principle* rule out from standards of rationality the ability to handle Toxin Puzzle-like situations in a decision-theoretically optimal way. And the optimal handling is that the agent forms the rewarded intention and then

[94] Bratman 1987, p.68: 'It is rational of the agent…to reconsider…just in case she thereby manifests reasonable habits of (non)reconsideration.'
[95] Ibid., p.70

reconsiders, so that they don't end up doing the bad thing that is intended – exactly what Bratman suggests is irrational.

For example, insomnia can be an occasion for Toxin Puzzle-like justification structures. Suppose an agent suffers from the following malady: a decision to work the next day would leave them unable to sleep, because they are terrified of the negative impact of the lack of sleep on work and so cannot relax enough to fall asleep; it would be better for their sleep if they decide the night before not to work the next day. If they make that decision and successfully sleep, then they know that it will make sense to revise that decision the next day, and work after all, since they have strong reason to work if they can. This exhibits the main characteristics of the Toxin Puzzle: an intention is justified because it aids sleep; the intention ceases to be justified once sleep is attained; what it is an intention to do is something it makes sense not actually to do if the intention is formed and has its beneficial effect[96].

Another instance is presented by Heuer:

> The New Date. Paul has arranged to take a new love interest to the cinema tomorrow night. This date makes him really nervous and jittery –and things are likely to deteriorate until tomorrow night. If he knew that his good friend Ellie intended to come to the cinema as well, he would feel a lot calmer. Ellie knows this, but she has no interest in seeing the film, and she doesn't believe that her actual presence would do Paul any good. It may be awkward or, at best, it would be irrelevant. So she has no reason to go to cinema tomorrow. But she has a reason to intend to go, because it would help Paul to calm down now.[97]

These situations are entirely ordinary and believable. If the question of how it is rational to behave is subordinate to the normality of such situations, as on Bratman's view, then the optimal handling of such situations must be part of reasonable standards of rationality. The normality restriction cannot adequately ground a totally general prohibition on rational reconsideration on intentions, one that applies even in Toxin Puzzle-like situations. Bratman's

[96] A different insomniac case, one that concerns an insomniac's insomnia-causing beliefs, is discussed in Harman 1976; it is taken further in the direction of intention by Bratman 1991, though not quite as far as the version suggested here. The case is unfortunately modelled on my own personal experience.
[97] Heuer 2018, section 6

conception of the foundations of practical rationality do not seem to bear out the rationality of nonreconsideration of intentions in its full generality.

There are two ways of taking this point further. The first is to suggest that on Bratman's understanding of the basis of the rational nonreconsideration requirement, the passage of time in the Toxin Puzzle ought to count as a problem for the intention formed at midnight to drink the toxin. That is, it ought to be make reconsideration rational in Toxin Puzzle-like situations, since a habit of reconsideration in such situations would be beneficial. If this is right, then the rational nonreconsideration requirement is not adequately grounded in the first place, so it certainly can't serve in a diagnosis of the irrationality of intending to drink the toxin.

The second way to take this point is that the rational nonreconsideration requirement ought to be itself suspended in Toxin Puzzle-like situations. That is, even if the passage of time does not count as creating a problem for the intention to drink, it just is a reasonable habit to reconsider in Toxin Puzzle-like situations, so that agents ought to reconsider even though there is no 'problem', in the technical sense, for their prior plan. Again, if this is right, then the requirement won't help to explain the irrationality of intending to drink the toxin.

However, we might move away from Bratman's own very particular and contestable understanding of the basis of the rational nonreconsideration requirement, and towards an alternative foundation. A separate and perhaps more natural way to understand the nonreconsideration requirement is as a reflection of the rationality of not re-opening completed reasoning unless there is some positive reason to suspect that that reasoning was not adequate i.e. that it relied on a false premise or on a premise that has become false. That is, one ought to reconsider one's intentions if the reasons one has for one's intentions relevantly change.

However, the problem with this should now be obvious. To rely on this in understanding the irrationality of intending to drink the toxin is to presuppose an account of what reasons there are to have intentions. Our interest is in a rational requirement that has an independent basis – so, for the sake of

denying premise 4), we have no interest in such an account (even if it happens to be correct in its own right).

As an illustration, consider a different version of the rational nonreconsideration requirement, one examined in the last chapter: Pink's suggestion that a rational intention leaves the agent able to act rationally (by their own lights). Combined with the assumption we are conceding, that it is not rational to drink the toxin when the time comes, it follows that it is not rational to intend at midnight to drink it. Pink's requirement is distinct from the non-reconsideration requirement; it rather asserts a more basic and primitive rational connection between intention and action. But as we have seen, Pink understands this requirement as grounded ultimately on a conception of decision itself as governed, in its applicable reasons, by the primacy of action, so that the job of the decision is to get the agent to perform appropriate actions.

This account *itself* rules out any creation by the financial reward of reason for the agent in the Toxin Puzzle to intend to drink the toxin, and so (even if it is correct) the rational requirement on the agent not to intend to drink the toxin cannot explain *why* the financial reward does not create reason for them to so intend. Rather, it depends on exactly that fact.

So we must consider alternative kinds of rational requirement that could interfere with the generation of state-given reasons in the Toxin Puzzle – but ones whose philosophical grounding does not rest in a conception of legitimate reasons for intention. In what follows I will consider the notion of aiming – what it takes to aim at an action. Since to intend an action is also to aim at performing it, an agent cannot rationally intend if they cannot have a coherent self-conception of themselves as aiming at their object. This can be exploited to create rational requirements on intention – as we will see, ones that can help us in explaining why the financial reward for the intention does not create reason to have it.

Rationality and ability: Argument 2

A second important argument against the rationality of intending to drink the toxin is provided by Kavka himself. The crux of this argument is contained in the following passage:

> we can explain your difficulty in earning a fortune: you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act. And you have (or will have when the time comes) no reason to drink the toxin, and a very good reason not to, for it will make you quite sick for a day.[98]

On one possible reading of this argument, it is baldly asserting what we have already rejected: that practical deliberation just is essentially concerned only with what the agent has reason to do, so the agent 'cannot', presumably in the sense of inability to do so rationally, intend something contrary to what those reasons justify.

However, I will briefly consider an alternative reading. (This reading involves some extrapolation, but the point here is to consider various arguments that *might* be made, not what Kavka himself thought).

On the reading I shall consider, it is impossible for rational agents to intend to do what they know they have good reason not to do. That is, rational agents are *unable* to intend to do what they know they have good reason not to do. But why might this be?

Here is one explanation. Rational agents, firstly, cannot rationally intend to do what they know they cannot do. It is plausibly a necessary condition on rational intention that it is possible for the agent to do what they intend. A rational agent (in normal conditions) cannot intend to do a high jump over Big Ben. Then, a second assumption: that rational agents cannot do what they know they have good reason not to do. (And, as a corollary, they cannot have a present-directed intention to do what they know they have good reason not to do.) It follows that a rational agent cannot drink the toxin, since they know they have decisive reason not to.

[98] Kavka 1983, p.35

Since agents cannot rationally intend to do what they cannot do, and they cannot drink the toxin, it follows that they cannot rationally intend to drink the toxin. To re-iterate, the key premises here are A) the irrationality of intending against one's ability, and B) the inability in rational agents to do, or have a present-directed intention to do, what it is irrational to do.

If the agent in a Toxin Puzzle expected to be less than fully rational when the time came to drink the toxin, they could perhaps get around this argument. Since they do not expect to be rational, they do not necessarily expect to be unable to drink the toxin. So, as far as this argument goes, there is no reason why that agent should not intend to drink the toxin. This is a curious implication.

The only premise that can reasonably be questioned is premise B): the claim that rational agents are unable to do what they know they have good reason not to do. However, the sense in which this seems true is just the idea that rational agents cannot *rationally* do what they know they have good reason not to do – and accordingly, that insofar as they are rational, they *won't*. For this premise to connect with premise 1), the irrationality of intending against one's ability, premise B) would have to imply that rational agents are literally unable to do what they know they have good reason not to do. It would have to amount, in effect, to the following claim: that rational agents are psychologically unable to act contrary to their reasons.

If this were true, it would follow that to intend to drink the toxin is to intend to do something that one knows one won't be able to do (given that one expects to be rational). And that in turn would be a good argument for the irrationality of intending to drink the toxin. But this is not an implication of the plausible reading of premise B). The inability to do something rationally does not imply an inability to do it *simpliciter* – not even if the agent is rational and cannot avoid being rational.

All sorts of strange consequences would follow if it were true that rational agents are not just rationally unable, but unable *simpliciter*, to act irrationally. For instance, we would lose the ability to make sense of a rational agent's choosing the rational course of action from among their options. If a rational

agent is unable to do anything other than act rationally, then it follows that the irrational courses of action are not options for them at all. At most, they only seem to be options, but since the agent is unable to do them, they cannot be intelligibly considered as something possibly to choose. Making sense of a rational agent's choice requires taking them to be able to pursue various options, but nonetheless as opting in favour of one in particular. The other, irrational, options cannot be excluded from deliberation as possible things to do. But this is what would be required if they are to be impossible in the way that prohibits them from being objects of rational intention.

Intending to drink the toxin is not like intending to jump to the moon. Even if it is irrational, it is not irrational in the same way. It relates to something that one *can* do, even if one is rational. A rational agent *can* drink the toxin, but, we are assuming, won't[99]. Their rationality does not *prevent* them from carrying through to its completion an intention to drink the toxin. Rather, the fact that an agent won't drink the toxin, and that they will abandon any prior intention to drink the toxin, is what makes them rational.

Intention and certainty of non-performance: argument 3

Despite the failure of the ability argument, it seems to get something right. It understands that the crucial difficulty in forming the intention to drink is the fact that the agent, in forming their intention, must look ahead to what it makes sense to do at the time of action. Understanding that it makes sense, when the time comes, to refuse to drink the toxin, this exerts some sort of effect on their ability to intend beforehand to that very thing – the very thing which they think it will make sense not to do.

Alongside the restriction on rational intention that it cannot be directed at an impossible object, there is a related restriction that proves more useful: that rational intention cannot be directed at an action which the agent is certain

[99] Bracketing, again, views such as Gauthier's on which this is in fact rational.

they will not do[100]. This latter restriction entails the former and, I shall now suggest, is appropriately seen as its explanation.

The obvious rationale that helps to explain why agents cannot rationally intend to do something they are unable to do also justifies the broader restriction that agents cannot rationally intend to do something that they are certain that they won't do. Intention seems to be constitutively connected to the agent's planning – to enabling the agent to undertake preparations so that eventually they will do what they intend. No amount of preparations can enable the agent to do what they will be unable to do; that's why it makes sense that intention cannot be rationally directed at an impossible object. For the agent to intend to do something impossible would be for them to be oriented towards preparing to do an impossible thing. Under that description, the agent is only wasting their time.

This point also helps to explain why intention cannot be rationally directed towards actions the agent is certain they won't do. Nothing the agent does will result in the occurrence of an event that is certain not to occur. If they intend to do something they are certain they won't do, then what they do out of that intention won't achieve its aim. These points run parallel to those that apply to intending an impossible object. Under the description, 'preparing to do something the agent is certain not to do' the agent is only wasting their time.

The former restriction, that rational intention cannot be directed at an impossible object, is an implication of the latter restriction, that rational intention cannot be directed at something certain not to occur, since anything (known to be) impossible is certain not to occur. This suggests that it is the latter restriction that is the fundamental one, and that explains the status of the former restriction on ability. In normal circumstances, the former restriction is the expression of the latter: normally, the reason why an action is certain not to occur is because it is impossible for it to occur. We exclude from our range of options only what we know is impossible. That is why it is tempting to treat the restriction of rational intention to possible objects as

[100] The possibility of this constraint is also mentioned in Levy 2009, fn. 1.

basic and intuitive – more so than the restriction of rational intention to objects whose non-occurrence is not certain.

The point that rational intention cannot be directed at something certain not to occur bears in an interesting way on the Toxin Puzzle. Unlike normal choice situations, in which the impossibility of an action is why it is certain not to occur, the drinking of the toxin is certain not to occur even though it is possible for the agent to do. Drinking the toxin is not excluded from the agent's range of options. Rather, it is something that they are certain they won't do.

The brief version of the argument is as follows. Provided that the agent is certain that they will be rational, they are certain that they won't choose to drink the toxin. They are certain that they won't do something that it is possible for them to do. If intentions cannot be rationally directed at events certain not to occur, it follows that an agent cannot intend to drink the toxin. They cannot intend to drink the toxin when they are certain that they won't drink the toxin[101].

Complexities arise when we consider to what extent the agent is certain that, if they are rational, they won't drink the toxin. If the agent does form an intention to drink, and does not reconsider their intention, then they will drink the toxin. So if they are to be certain that they won't drink, this means that they must be certain that, even if they form the intention now to drink the toxin, they will reconsider this intention after they receive the financial reward for forming that intention.

In this vein, some argue that an agent in a Toxin Puzzle-like situation is rationally required to do exactly this: to form their intention to drink the toxin in such a way that they know in advance that they will not reconsider this intention[102]. With this knowledge, they cannot be certain that they will fail to drink the toxin.

---

[101] This possibility is envisaged in Shah 2008, p.16-7 but not refuted there, despite Shah's ambition to use the Toxin Puzzle to motivate his view that practical deliberation ought to be governed only by considerations bearing on the worthwhileness of the action.
[102] As argued in Holton 2004.

Agents who are capable of such super-resolutions therefore lie beyond the reach of this argument. It is, it seems, genuinely rational for them to form the super-resolution to drink the toxin. Nonetheless, for agents who are not capable of such super-resolutions, they can be practically certain that if they are rational they will reconsider their intention to drink the toxin once they receive the reward ('practically' certain because it is still possible that some contingent event could interfere with their reconsideration). Such agents are certain that they won't drink the toxin. And so they cannot rationally intend to drink the toxin, being certain that they won't do it even if they now intend it.

It is important to note that for this argument against the rationality of intending to drink the toxin to work, an underlying conception of rationality similar to Bratman's must be avoided. As discussed above, on Bratman's conception of rationality, rational behaviour is behaviour that contributes to the agent's long-term interests in normal conditions. Assuming, as argued above, that Toxin Puzzle-like situations are within the bounds of normality, it follows that rational behaviour in such situations requires intending and then reconsidering one's intention post-reward. This is in tension with my insistence that intentions can never be rationally directed to what is certain not to occur. So a conception of rationality such as Bratman's is not consistent with this sort of argument.

Returning to the Anti-Desirability Argument

I have suggested that we can block the conclusion of the Anti-Desirability Argument by suggesting that the failure of the desirability of the toxin-drinking intention to create a reason to adopt that intention could be unique to Toxin Puzzle-like situations – depending, in particular on the temporal structure of the reward and the fact that the action involved both is undesirable and has nothing going for it. That the irrationality of a rewarded intention could be unique to such situations would be true if the explanation of its failure to justify was that any such intention would be irrational. Put more intuitively, it wouldn't make sense to intend to drink the toxin, and that's why

that intention cannot be justified, or rationalized, by the reward offered for so intending.

The points made above help to substantiate this line of thought. A rational requirement on intentions is that the agent does not intend something they are certain will not occur. In application to the Toxin Puzzle, in the absence of deliberation-avoiding devices, the agent is certain that they won't drink the toxin, and so they cannot rationally intend to drink it. This is the independent rational requirement on intentions we need that can be understood prior to considering what reasons there are to have that intention. It is because the agent can be certain that they won't drink the toxin that it doesn't make sense for them to intend, or plan, to drink the toxin. To do so would involve them in an incoherent self-conception: as aiming to do something that they expected not to do[103]. Therefore, nothing could justify that intention for them, even if they would be better off if they had it.

This diagnosis of the irrationality of that intention also helps to explain something noticed earlier: why it is natural in the case of the Toxin Puzzle to take the acceptability of drinking the toxin itself to be foremost in the agent's practical deliberation, even though this intuition as to the primacy of action is one we have suggested does not apply wherever the agent considers what to intend.

The reason is that, in order to settle whether or not it is certain that they won't do what they intend, the agent must look ahead at how things will strike them at the time of the action itself. In thinking through what it makes sense to do in a Toxin-Puzzle-like situation, this rational requirement on intentions makes it natural to evaluate one's intention with reference simply to whether it will make sense, at the time of action, to perform the intended action. It is because the rational action considered at the time of action is not the action that the desirable intention aims at that that intention is neither rational nor justified.

[103] Bratman himself also makes a self-conception argument for the irrationality of intending to drink the toxin in a separate piece (Bratman, 'Following Through with One's Plans: Reply to David Gauthier', in Danielson (ed.) 1998, ch.4 p.57-8): 'I do not think that such a pragmatic, two-tier approach exhausts the subject…We can also appeal directly to a kind of incoherence involved in intending and attempting to A while knowing one cannot A. If I am at all reflective, I cannot coherently see what I am doing as executing an intention to do what I know I cannot do.'

This creates the appearance that intentions are generally subject in practical deliberation to the worthwhileness of their actions. This appearance is a product of the Toxin Puzzle's unique set-up and does not support general conclusions. By explaining this appearance, we avoid a theoretically unsatisfying dismissal of the intuitions that support it – because the irrationality of intending to drink the toxin really does depend centrally on the irrationality of drinking it.

What makes for the irrationality of intending to drink the toxin does not generalize, in this diagnosis, to non-Toxin Puzzle-like situations. The Toxin Puzzle turns on the fact that the intention ceases to be desirable in advance of the time of the intended action. This feature is absent from our cases discussed in the previous chapter.

Conclusion

Our initial survey suggested that the only serious option for dealing with the Toxin Puzzle, given that our overall thesis is that state-given reasons exist, is to simply accept that it is rational to intend to drink the toxin. Our detailed examination has suggested that it is after all irrational to intend to drink the toxin, but also that there is a second option for dealing with the puzzle: the possibility that state-given reasons fail to exist specifically in Toxin Puzzle-like situations. This option relies on the fact that intending to drink the toxin breaks a rational requirement on intention understood independently of theories of reasons for intention.

However, this still leaves us with a mildly unfortunate situation. Philosophers who conceive of rationality as reasons-responsiveness, or as something similar, will not accept this way of blocking the general conclusion that state-given reasons do not exist. They are likely to accept that the Toxin Puzzle forces the theorist to choose either between the thesis that it is rational to intend to drink the toxin, or the thesis that state-given reasons for intention do not exist. The implausibility of the former will then lend support to the latter, even though the latter is, if my arguments in the previous chapter are correct,

itself vastly implausible. The Toxin Puzzle is useful in forcing us to confront the limitations of this way of conceiving of rationality.

4. On action as the conclusion of deliberation, and the significance of this for intention

The previous chapters examined the topic of state-given reasons for intention from the perspective of a) comparisons with belief, b) intuitions about sensible intentions in various cases, and c) a decision-theoretic puzzle – the Toxin Puzzle. Although at each point a defence was made of the genuineness of state-given reasons for intention, each of these lines of inquiry is indirect, staying away from fundamental considerations concerning what intention is and does. This next part of the thesis (chapters 4 through 6) tackles state-given reasons for intention from this deeper angle.

In chapter 1, I specific the usual conception of what sort of reasons intentions rationally answer to: the conception on which they answer only to object-related reasons, i.e. reasons bearing on the actions one intends and what makes those actions worth performing. This contrasts with an alternative possible conception on which intentions should respond to reasons bearing on the wisdom of intending itself. On the former view, intentions are rationally acquired when one rationally arrives at an answer to a question of the form 'What to do?'; on the latter, intentions are rationally acquired when one rationally arrives at an answer to a question of the form: 'What to intend?'. (On certain further views, the answer to this latter question is itself determined by the answer to the former). Object-related reasons are, clearly, crucial for practical deliberations out of which intentions arise; an adequate view of intention that admits state-given reasons for intention must explain how object-related reasons get into the picture, and this task is carried out in chapter 6. For now, however, we need to look in more detail at the picture we reject – that intention is rationally controlled only by object-related reasons, those that bear on the wisdom of doing what it is an intention to do. We must look at this restrictive conception from the perspective of thinking about the nature of intention.

This is the background to the argument made in this chapter, in which I suggest that doing justice to the idea that action (not intention) is the

conclusion of deliberation requires rejecting the idea that intention is the conclusion of practical deliberation focused on the question 'What to do?'. I don't argue *that* action is the conclusion of deliberation: doing so would be a vast undertaking and would fall outside the scope of the issues examined in this thesis. The suggestion is, rather, that *if* that is true then it is wrong to conceive of intention as the proper conclusion of thinking about what to do. In essence, the argument is that if intention is that sort of conclusion, then one can only rationally arrive at a future-directed intention once deliberation about what to do is concluded: once one has reached a conclusion on what to do. But then no deliberation is left to be done in between the formation of future-directed intention and the beginning of action itself, so there is no deliberation for action itself to conclude. By *modus tollens*, intentions are not conclusions of deliberations focused on the question 'What to do?'.

The idea of intention as a conclusion to deliberation focused on 'What to do?' is not equivalent to the idea that intention is rationally controlled only by object-related reasons. As I shall suggest, even if intention is not grounded on an answer to that question, perhaps it expresses an attitude towards that question nonetheless, an attitude itself grounded in what the agent thinks of their reasons for action. So the idea that intentions are not the conclusions of deliberations focused on 'What to do?' does not itself *entail* that state-given reasons apply to intentions. But it does help to undermine the most natural version of the idea that intentions are rationally controlled only by perceived reasons for action, which is just the idea that intentions are products of what the agent thinks the answer to 'What to do?' is. This is Shah's and Hieronymi's idea, and it is one that the considerations mounted here help to refute.

Instead of drawing a comparison between intention and belief, this chapter helps to motivate an alternative scheme: where intention is acquired as the proper answer to the question 'What to intend?', action is performed as the proper answer to the question 'What to do?' (and belief as the answer to questions about what is the case). Once we admit the idea of 'what to intend?' as distinct from 'what to do?', allowing intention to be the answer to the former introduces the question of what sort of thing an answer to the latter is.

To construe intention as the answer to both 'What to intend?' *and* 'What to do?', as the conclusion of deliberation on each question, would invite the objection that the two questions are not really being treated as distinct. This chapter pre-emptively answers that objection.

Practical Questions

Whenever an agent deliberates, there is some question they are deliberating on, with a view to answering. Practical deliberation is, roughly, deliberating on how to act, or what to do: the question here is something like 'how shall I act?', 'what shall I do?' or 'what ought I to do?' and practical deliberation concludes when that question is answered. Likewise, doxastic deliberation, or doxastic inquiry, takes as its aim the answering of some specific question and concludes when those questions have either been answered or judged unanswerable.

This is a substantial, and contestable, assumption. Instead of conceiving deliberation as the answering of a question, we could conceive of it more minimally as a kind of psychological process, and characterize it in terms appropriate to psychological process. Perhaps the most well-known view of this kind is in Thomas Hobbes:

> Deliberation— When in the mind of man, Appetites and Aversions, Hopes and Feares, concerning one and the same thing, arise alternately… the whole sum of Desires, Aversions, Hopes and Feares, continued till the thing be either done, or thought impossible, is that we call DELIBERATION.[104]

All sides can agree that deliberation is a process that can be stopped and restarted or resumed. It may terminate without being concluded – such as when I'm too tired to continue, and subsequently forget about the whole thing. But when it concludes, it terminates. Hobbes forbears from tying deliberation to the answering of any question. But having noted the deniability of this assumption, I will continue to accept it in this chapter. I assume that deliberation concludes with the answering of the deliberative question: the

[104] Hobbes 1994, ch. VI

event of reaching of an answer to the deliberative question just is the event of deliberation's concluding.

Accordingly, once a question has been answered, then to deliberate more on that question is necessarily to reconsider one's answer. The detective who inquires further into a death after it was judged unsolvable is, necessarily, reconsidering whether it is really unsolvable. The relation agents bear to the questions they have views on is single: they either take themselves to have an answer to the question, or they don't, and if they do purport to have an answer, then whatever they give by way of an answer is, collectively, uniquely and singly their answer.

This chapter discusses how this bears on the practical. As mentioned, practical deliberation addresses some deliberative question along the general lines of 'what shall I do?' or 'what ought I to do?'. This chapter considers when these questions are closed, and when they are reopened, with a view to determining what the questions *are*, and when they are pertinent, and what counts as an answer.

There is an argumentative move here that this chapter hinges on. If we want to know what the practical questions are, then we can consider when these questions are closed and when they are re-opened. If we can be sure that a specific question is closed and not re-opened, then we know that whatever deliberation occurs later than its closing is deliberation directed on a different question i.e. a question with a different content. That is, the usability of this method is prior to our knowing what the question's content is.

To answer a doxastic question is to acquire a relevant belief. Insofar as I continue to hold that belief, I continue to be disposed to answer the relevant question in the same way, viz. the way encapsulated in my belief. For example, if today I acquire the belief that the government will fall approximately within the next six months, and I retain that belief throughout tomorrow and the day after, then throughout that time I am disposed to answer the relevant question(s) in the same way: 'when will the government fall?' by 'in the next six months' and 'will the government fall in the next six months?' by 'yes'.

This is just part of what it means, I think, to attribute a continually persisting mental state of belief. It means that there is some range of questions which we are continually disposed to answer in the same way: whatever are our answers to these questions, we don't reconsider them. Though there may be complications here, I suggest that the above is at least tied conceptually to belief and is not merely a normal feature of cases in which we ascribe a relevant belief.

Perhaps the same is true of intention. (In this chapter, I will use 'intention' to refer, specifically, to future-directed intention unless stated otherwise.) On such a view, to attribute a continually persisting intention is to say that there is some range of questions which we are continually disposed to answer in the same way, whose answers we don't reconsider. This will be contentious: it incorporates the assumption that intention is, or at least involves, an attitude towards a certain proposition.

The purpose of this chapter is *inter alia* to explore the usefulness of this assumption. But the assumption is initially plausible. It seems that, at least normally, if an agent intends to make themselves a cup of tea, then there is some range of questions to which they hold certain answers: 'what shall I do now?' with 'make a cuppa' and 'when shall I make a cuppa?' with 'now'.

There are different candidates for what this range of questions, and their answers, might be. But to say that intending involves the settling of some relevant range of questions is not to say that every intention must involve the settling of the same kind of question. This is part of what's at issue in the debates regarding the nature of intention, but it is important to disentangle it from the underlying assumption that as long as an agent retains an intention, they are treating some question or other as answered.

Views on what it means to retain an intention have implications for which deliberative questions are closed and which left open insofar as that intention is retained. The content of these implications depends on a bridging principle that specifies the kind of propositions, if any, to which having an intention commits us.

Reconnecting this to the aim of the chapter: if we want to understand which are the questions that practical deliberation answers, then we need to understand two things. Firstly, which are the questions that are closed when one has an intention: these questions are settled when the agent adopts an intention through practical deliberation. Secondly, we need to understand which are the practical questions that might be unanswered even as the agent has and retains an intention. The questions that having an intention does not require treating as answered are the questions that practical deliberation might answer even after an agent has, through practical deliberation, arrived at an intention. So if we know that practical deliberation occurs at a certain time t, and we know that the agent must be treating a certain question as closed at t owing to their having retained a relevant intention, it follows that any practical deliberation at t cannot be answering that question.

This chapter avoids making a certain general assumption: I do not assume that which questions an agent has answered is, even in a rational agent, necessarily closed under consistency. So, for any two questions Q1 and Q2, and answers A1 and A2, even if Q1's being answered with A1 is consistent with Q2's being answered only with A2, it does not follow that if the agent holds A1 as the answer to Q1 then they hold A2 as the answer to Q2. So when I suggest that deliberation permits treating certain questions as not yet answered, I am not suggesting that when the agent deliberates there is necessarily more than one answer to those questions that the agent could consistently adopt, given their other commitments.

As mentioned earlier, there are multiple kinds of statement that have been identified as the proposition one holds true insofar as one has an intention. 'I ought all-things-considered to φ' and 'I shall φ' are perhaps the simplest. Some of them are rather abstruse[105]. Since I am trying to judge the validity of the whole approach, I will not commit to any specific form of the question. Instead, throughout this thesis, I use a general formulation that is only barely acceptable in English: 'what to do?' and, correspondingly, 'to φ'. I intend this

[105] I have in mind Setiya's formulation of what it is to act for a reason, in his 2007. Though not intended as a general account of intention, it is intended as an account of acting intentionally and can naturally be extended to cover future-directed intention.

not to contrast with the above alternatives, but rather to be what they purport to capture. The aim here is also to leave open the possibility that none of those above alternatives capture the question which intention involves answering. At this stage it is a placeholder formulation.

The Puzzle

It is now time for a brief statement of the puzzle I wish to critically explore. Suppose that future-directed intention involves the closing of some relevant range of questions along the general lines of 'how to act (in the future)?'. And suppose the agent retains this intention up until the time of action. It follows that that question, how to act (at the relevant time), is not open at the time of action. So it is not answered at the time of action by anything like practical deliberation, and consequently, is not answered by a practical deliberation which concludes *in an action*. So we have a quick argumentative route to the idea that action is not interestingly the conclusion of deliberation: never the conclusion of deliberation when there is a prior intention, and where there is not a prior intention, the conclusion of deliberation only by grace of the absence of prior intention.

If there is something interesting in the idea of action as a conclusion of deliberation, then it must be a kind of deliberation that in principle could conclude *only in an action*. This is a minimal requirement on an interesting version of this thesis. Action's place as the conclusion of deliberation cannot hold merely in virtue of when the deliberation happens to occur.

This is a nonstandard question in the literature on the conclusion of practical reasoning, to the extent that any questions in that literature are taken to define the issue. I am concerned with this issue because it seems to me to afflict several of the proposals that have been put forward.

For instance, Phillip Clark claims that what Aristotle meant by this idea was that intention is the product of practical reasoning and that intention has action as its content; action is the conclusion of deliberation in the sense of being the content of the final intention-product of good practical reasoning. 'On the face

of it, what I intend when I intend to make a cloak is an action, namely the act of making a cloak'[106]. What *motivates* Clark is a desire to avoid the objection that 'if a conclusion is drawn but no action occurs, then the conclusion drawn is not an action'[107]. He spells this out with Kenny's example:

> Suppose I think for a while about how to use some money I've inherited, and I decide to buy a piano as soon as the quarter is over. Sadly, I do not live to the end of the quarter, and consequently never buy the piano. Kenny notes that my failure to do the action I've decided to do hardly renders my reasoning inconclusive. I do reach a conclusion about how to use the money, and I reach it well before the scheduled time of action.

This is just the puzzle I am concerned with, that there is no sense in which the event of action is the event of deliberation's concluding, just because there is no question left for reasoning to answer at the time of action[108]. If this puzzle could be tackled in its own right, then the idea of action as a conclusion of deliberation wouldn't have to be interpreted to make it consistent with it.

Some writers draw an analogy between action as the conclusion of practical deliberation and belief as the conclusion of doxastic deliberation. Dancy rejects the idea that the relation of action to deliberation is properly described as a concluding relation, because he thinks it requires a nonsensical idea of inferring action from one's reasons. But he then insists that minus that the whole issue has an 'easy answer'[109], just consisting in the fact that

> when an agent deliberates well and then acts accordingly, the action done is of the sort most favoured by the considerations rehearsed, taken as a whole—just as when an agent reasons well and then believes accordingly, the belief formed (the believing, that is, not the thing believed) is of the sort most favoured by the considerations rehearsed, again taken as a whole.

---

[106] Clark 2001, p.501
[107] Ibid., p.483.
[108] Given the questions-centric framework I am using, one could suggest that action answers the question 'What to do *now*?' whereas future-directed intention only answers the question 'What to do later?' – so the questions are distinct after all. This would involve treating the 'now' as contributing something different to the overall proposition than 'later' does so that 'I shall φ later' and 'I shall φ now' emerge as distinct propositions. I do not have a refutation of this suggestion, but it clearly connects to larger issues in temporal thought that are beyond the scope of this thesis.
[109] Dancy 2018, chapter 2.2

That is, on Dancy's view, action is exactly like belief with respect to deliberation just in that both are responses to the reasons rehearsed when deliberating.

But the central problem arises quite apart from what we think about the relation of conclusions to inferences. There cannot be deliberation, as I have suggested, unless there is some question that defines the deliberation that is being taken as unanswered when one deliberates, where the deliberation's purpose is to answer that question. What is the question being answered in practical deliberation? If it is 'what action do my reasons favour?', then that question is answered just as soon as one judges what one ought to do, and once that best judgment is reached, deliberation ends. And if it is 'what shall be my response to the reasons?' then that is answered, if it is answered at all, when one forms an intention, and that is where deliberation ends. And this opens the way for the initial puzzle I specified to get going.

This speaks directly to Dancy's concern, which is to sort out differences between practical and doxastic deliberation. For there is nothing like future-directed intention in the case of belief. To acquire a belief on some matter is to answer some question concerning that matter; the deliberation concludes then. When one acts, then according to the suggestion at the heart of our puzzle, the question 'how to act?' has *already* been answered. So an action cannot play the role in practical reasoning played by belief in doxastic reasoning, if that conception is correct. And if there is no more reasoning to be done, because all the relevant questions have been answered, then we lose our grip on how the event of action can have any relation to practical reasoning other than by being among its effects.

There is something right in Dancy's ambitions on the subject: action's status as the conclusion of deliberation, or near enough that, follows from the existence of reasons that are properly described as reasons to act. Dancy aims to respect the requirements on intelligibly describing things as reasons to act, and one of those requirements is that action sometimes gets to be a response to those reasons. But I suggest this requires more than Dancy's 'easy answer',

if this ambition is spelled out in terms of action's playing a relevant role in practical reasoning itself.

This disanalogy between practical and doxastic deliberation is also an issue in Fernandez' view. Fernandez asserts, in line with our own ambitions, that the event of acting is an event of concluding some practical deliberation. Hence concludings are extended happenings:

> The conclusion of practical reasoning is not a fully determinate particular. The act with which practical reasoning concludes or, rather, the act of concluding itself is the doing of the action, the action-in-progress. The reasoning has not reached its completion until the action has.[110]

But if action concludes reasoning, then what is the reasoning that leads to the adoption of an intention? In a footnote, Fernandez suggests that it is the same as that which results in action[111], as long as we suppose that 'intending or otherwise practically judging are already forms of acting'. If this is so, then intending is the first part of an agent's answering the practical question, with their performance of the relevant action a later part of their answer. And yet it seems puzzling that when one intends, there should be *no* question that the agent is treating as *already* settled or answered. For when I intend I must already be confident that what I intend is a good idea; so *some* questions, at least, have already been answered. But, if action together with its prior intention are the constituents of the event of deliberation's concluding, then there is still some question that is being answered *in* that concluding. Yet if this question is 'how to act?', then there is, at least, a *risk*, that this question will already have been answered in virtue of a prior intention. The risk is that having an intention involves treating as answered just those questions that are plausible candidates for the questions that deliberation that concludes in an action thereby answers. This is what our opening puzzle hinged on. These are complexities that just don't arise in the doxastic case with which Fernandez hopes to draw a parallel, and they ought not be skipped over. This isn't a true objection, but rather my attempt to say what exactly is so compelling in the tradition Fernandez rejects.

---

[110] Fernandez 2016, p.896
[111] Ibid., fn. 60

We can sharpen the puzzle by presenting the choice point in terms of a need to avoid an inconsistent conjunction of three propositions, all of which are intuitively attractive.

Proposition 1: Adopting an intention involves settling some determinate range of questions on how to act, where the range remains the same over time including the time of action.

Proposition 2: At the time of action there may exist an open practical question, how to act, which action itself settles (and this may exist even when there is a prior intention in play).

Proposition 3: It is possible to adopt an intention to act without the intended action thereby starting (this is sometimes true even in cases where, after one adopts that sort of prospective intention, there remains an open practical question which action itself settles).

The views briefly discussed in the previous section aimed to respect the idea that action is the conclusion of deliberation, but ended up making it puzzling just how we should read that idea. Clark and Dancy don't appear to endorse proposition 2. Fernandez is more naturally construed as rejecting both propositions 1 and 3 in order to protect 2. But then a question arises for his view: given that we only need to deny *either* 1 or 3 in order to protect 2 from this inconsistency, why deny both?

The reason is that on his view, even once action has already started, one's reasoning is not yet complete; it only ends when the action does. So if practical reasoning requires the openness of a relevant practical question, the relevant practical question must remain open right up until the action is finished. So it cannot be that adopting an intention closes that question; it can only be that, at most, adopting an intention *initiates* the closing of that question that is the action, on Fernandez' view. Hence, at least, his need to reject 1 as well as 3.

This idea incorporates the strong assumption that the same piece of reasoning is in play the whole way through, and this seems to me a genuine presupposition of Fernandez' view that must be explored: that if I intend to

rob a bank, plan the heist, turn up, threaten everyone, empty the cash registers and run out, that is all the same piece of reasoning issuing from my premise that I am getting rich quick. Fernandez' notion of a 'rational order of acts'[112] appears not to permit any distinctions between the components of higher-level actions: each part of the means is equally justified and explained by the end. It is to be wondered whether an alternative reasoning structure might cohere better with our intuition that when an agent intends then there must be some question, regarding how to act, that they are treating as already answered – that is, with proposition 1 – and whether, along with that, some relevant reasoning is already complete.

Proposition 3, equally, has much to be said for it. Korsgaard denies it:

> It is frequently argued that intentions must exist separately from actions because we often decide what we will do (and why) in advance of the time of action. I believe, however, that we begin implementing or enacting our decisions immediately, for once a decision is made, our movements must be planned so that it is possible to enact it, and that planning is itself part of the enacting of our decision.[113]

But I can intend to do something without yet planning anything. If an agent is an experienced bank robber, then they might intend to rob a bank sometime this week but leave the actual planning of it until the last minute. And still, for the few days between now and then, they have a genuine intention. The Fernandez conception insists that in intending the agent is already acting in some sense, but since the experienced bank robber does not do anything at all in the way of planning or preparation until the last minute, there is nothing at all they are doing, hence no way they are acting. So the way we naturally describe our intentions, at least, seems to conform to proposition 3 and not to its denial. It is difficult to see what the Fernandez/Korsgaard view adds besides a verbal decision to call intentions actions or action-parts. Even if we set aside this problem, a further one looms: if intentions are actions or action-parts, then why should we suppose that they are part of the very action that is intended? This is what is required by the denial of proposition 3, and yet there is no obvious way to settle this issue, or even to determine what might be

[112] Ibid., p.886
[113] Korsgaard 2008, fn. 28.

meant. For example, one might suggest that even in adopting a partial plan for the future, one is already 'doing' something in ruling out other possible partial plans and this counts as performing some kind of action; but this pushes back the question to whether this really counts as an action as opposed to some kind of practical stance or, as I shall suggest in following chapters, an orientation of agency.

So it seems that we should think of practical reasoning in a way that makes room for both proposition 3 and something like proposition 1: at least, we appear to naturally describe our actions and intentions in a way which conforms to both 3 and something like 1. So this turns us back to proposition 2: perhaps that is the one to drop? More needs to be said in direct defence of proposition 2. There might be many sources of support for this view, but what I want to discuss is one source that is more relevant to my broader philosophical aims.

Suppose that there are some reasons to act that spring into existence only after the formation of both a judgment on what one ought to do, and an intention to do that. Then if there are those reasons to act, and the agent can act, and the agent is in a position to know about these reasons, then it seems that the agent should be able to act on these reasons. And if there are reasons for action that the agent can act on, then it seems that they must be capable of thinking about those reasons. And if they can think about reasons to act, and through that thinking come to act on those reasons, then it seems that the agent must be capable of practical deliberation. And if practical deliberation happens in these circumstances in which both a judgment on what one ought to do, and an intention to do that, are already in place, then action itself must be the conclusion of such a deliberation.

This may seem quite a lot to derive from the single initial premise that there could be some reasons to act that pertain only post-best judgment and post-intention. But there is something intuitive here. We call something a 'reason to act' when, by taking it up in deliberation, we become motivated to act, and capable of acting on it. So if there are reasons to act which pertain only post-

best judgment and post-intention, then deliberation post-best judgment and post-intention is possible too.

This isn't a general thesis about reasons. Perhaps there are some things that there may be reasons for and against, but where they are reasons that we cannot deliberate on. Reasons for emotion may be of this kind. It is plausible that there are reasons for feeling certain emotions: for example, that someone barged into me while rushing around, is a reason to feel resentful. But what connects an agent's feelings to their reasons isn't deliberation: though they can think about their reasons for and against emotion, deliberating on their merits, reasons for emotion are not productive *solely in virtue of* that deliberation – rather, if anything, despite it. Still, it is plausible when applied to reasons for action. It seems that if we can act for a particular reason at all, we must be in principle capable of coming through deliberation to act for that reason. Deliberation in this case cannot be categorically screened out as it is in the case of emotion. That is a possible view, but just quite a puzzling one: it would be strange if a reason to act could never, even in principle, be taken up in a deliberation with the result of motivating the agent to do something they weren't already motivated to do.

This general motivation for thinking of action as the conclusion of some deliberation has a lot to do with deliberation itself. If action *were* the conclusion of some deliberations, we would expect this to show up in the philosophy of deliberation itself, and not just in the philosophy of action. So part of the aim of this chapter is to substantiate this.

Just by being motivated by mundane thoughts about reasons, then, we are led to posit the possibility of action as the conclusion of some deliberation. So the initial puzzle we started out with is compelling: future-directed intention seems to involve treating certain practical questions as answered, ergo when one acts from a future-directed intention, the action itself is not the answering of those questions. Yet the substantiation of the above reflections will persuade us that even in these cases where there is a future-directed intention, we ought sometimes to allow for action to be the conclusion of its own

possible deliberation, and the answer to its own practical question. If this is right, then we must deny either proposition 1 or 3.

There is also another apparent strategy. One could defuse the inconsistency straightaway by weakening proposition 2 to the claim that the practical question action answers is not among the questions covered under the placeholder question 'how to act?'. Since, as I specified earlier, this is a general covering question, carrying through this proposal would involve sidestepping the resources of natural language in the identification of practical questions. But it is not clear that this solves the puzzle; rather it marks the place where a solution would need to fit. Naively, it doesn't seem that there is a question distinct from 'how to act?', or any of its subspecies, that is answered when I act. It seems rather, that once an agent finds the answer to that question, how to act, then their relevant reasoning is complete. So if we are to solve the puzzle by separating off the questions that action answers from the questions that intention answers, we will have to do this by weakening proposition 1 instead, and restricting the range of questions that intention involves having answered.

So we should look for a way to deny 1. I suggest that we do this by denying that if there are questions that intending constitutively involves treating as answered, and if there is a question that action itself sometimes answers, then they are necessarily the same question. So we should allow that there may be practical deliberation that concludes in an intention and thereby answers some relevant practical question, and we should also allow for there to be a second practical deliberation that concludes in an action and therefore must answer some relevant practical question. Though the relevant practical questions must occur at these points, we do not yet know what these questions are. Hence, the rational possibility of this view hinges on a plausible identification of the practical questions which each of these deliberations concludes by answering, and on the plausibility of the view that those questions are different.

Now as briefly discussed earlier, it does not naively seem that there is some question other than 'how to act?', or one of its familiar relatives, that is

answered when an agent acts. So if we hope to substantiate the above view, we will require that it is *this* question, 'how to act?' that is left unanswered by an intention, and that is answered only when an agent acts. So this view would deny that intention constitutively involves treating the question 'how to act?' as answered. It consequently accepts that the possibility of future-directed intention is consistent with the possibility of the practical question 'how to act?' being unanswered.

Even where the agent has a future-directed intention, then, since the question 'how to act?' is unanswered, action may be the conclusion of deliberation focused on the question 'how to act?'. Now it seems strange that an agent should, for example, intend to resume work later in the evening and still casually not have a commitment on what they are going to do later. So there is considerable intuitive pressure to accept that if an agent intends, then the agent must indeed be treating the practical question 'how to act?' as answered – contrary to the aspirations of our way of denying 1.

Moreover, as soon as the agent does answer the question 'how to act?' with a verbal formulation such as 'to φ', without the φ-ing starting, then the most they have is an intention. If that's what happens, then the answering of this question is the acquisition of an intention is the conclusion of the deliberation, and we have once more lost our grip on the idea of action as alone being the answer to this question and the conclusion of the corresponding deliberation.

So, if this way of denying proposition 1 is going to work, we need to deny that the question 'how to act?', understood correctly, *could* be answered with a verbal formulation. And we need to deny that such a verbal formulation is taken to answer the question when the agent has an intention.

We can connect this back to my earlier stipulation that we should not understand the question 'how to act' as, necessarily, being identical to one of the more familiar and construable questions such as 'how ought I to act?' or 'how shall I act?'. At least the former, and perhaps the latter as well, does have an answer that is verbally formulable in the absence of any action. Denying proposition 1, but holding onto the idea that action can be the conclusion of deliberation and that deliberation is the answering of an open

question, requires that we assign to action the function of answering a verbally formulable question, how to act, that is not answerable prior to the action.

Earlier I exploited the apparent existence of verbal answers to this very question, 'how to act?' – for example, 'what to do at six o'clock?' with 'have tea', thought before the time of action itself. We are forced now to treat the validity of those answers as illusory. The closest intelligible version of this is rather the following: we require that 'have tea' serve as the *presumptive* answer to 'what to do later?', where it is the action of having tea that is the actual answer. Analogously, before the presidential primaries, the candidate leading by a large margin in the polls is the presumptive nominee, and they remain the presumptive nominee even if there is no conceivable way they could lose. They are only the actual nominee once the primaries have completed and are officially designated the nominee. So 'who will be the nominee?' has only a presumptive answer before the completion of the process, even if we are certain that the answer will turn out to be true. The answering itself comes with the completion of the process.

Irrespective of that analogy, this proposal is clearly quite obscure. It is obscure how there could be something which is usefully characterized as a question, but which could not have as its answer something that falls short of action: a question that is verbally formulable, but whose answer is not. Nonetheless, this seems to be the best solution out of those I have considered so far.

Instead of delving further into the metaphysics of action, I will spend the rest of the chapter trying to provide independent motivation against the claim that if an agent has an intention to act then they have some answer to the question, 'how to act?' that is anything more than presumptive in a way that would foreclose further reasoning. The material will also serve to motivate further the view, mentioned above, that there are reasons which only pertain post-intention and post-best judgment: this is a significant motivation for accepting proposition 2.

The next section aims to substantiate this style of solution with reference to several test cases.

Reasons for action post-intention formation

What would be helpful at this stage is an independent proof of the possibility briefly adumbrated above: that an agent may, without compromising their rationality, intend to φ without yet treating 'to φ' as the answer to a relevant range of practical questions. If we were certain that that were possible, then we would have good reason to accept my solution to the inconsistent trio: that action may be the conclusion of deliberation when an agent fails, despite their intending to φ, to treat the practical question 'how to act' as answered – so proposition 1 is false. This section attempts to provide that proof.

I shall give a series of examples, and use a particular argumentative strategy to support the conclusion that they all instance this very state of affairs, in which an agent intends to φ but does not treat the practical question, 'how to act?', as answered. I shall discuss the argumentative strategy first, before putting it to use in the discussion of examples.

Each example is an example of deliberation that is supposed to instantiate the following relevant qualities: a) the agent is deliberating on whether to φ; b) the agent is deliberating on whether to φ because they intend to φ. It follows from a) that the agent is not treating 'to φ' as their answer to the question 'how to act?'. Ergo, they intend to φ if b) is true, yet they do not treat 'to φ' as their answer to the question 'how to act?'. *A fortiori*, intending to φ does not constitutively involve treating 'to φ' as one's answer to the question 'how to act?'. And since they have both an intention and a judgment about what they ought to do, if their deliberation concludes, it concludes in an action.

Sceptical readers will probably insist that b) cannot ever be true of a rational agent, or more strongly cannot be true of any agent, so they will read the examples differently to me. Perhaps they will deny that a) is true on some natural reading, or that the stronger claim, b), is true on some natural reading. These cases are complex enough that many ways of filling them out are

possible: I rely only on the idea that on some way of filling them out, a) and b) are true.

I have selected examples that evince some degree of neurosis. This is because the examples share the feature of the agent attempting, through deliberation, to head off incipient akrasia. Neurotics are especially practised at this: at handling their own recalcitrant motivation through a reparative kind of rational deliberation. So they are the best source of examples. I imagine readers with some experience of neurosis will find these sorts of situations familiar.

I have sketched the basic situation in each example initially, before moving on to fuller discussion. It is important that my fuller exposition be seen as one possible way of filling out the basic case, a way which suffices to prove the possibility of b) being true sometimes, without setting up an overly complex scenario from the outset.

Here follow the examples.

1. Someone may decide to bake a cake for someone as a way of showing the sincerity of their apology to them; when it comes to the time to bake, they feel too grumpy to want to do apology-related things, but after thinking about it, eventually motivate themselves to bake the cake with the spiteful thought: 'This'll make them feel bad for getting so mad at me'.

2. An intensely self-critical person might aim to do something nice for themselves and obtain a little relief by exercising, but then when the time comes not be in the right place to carry out this benevolent aim. So instead they turn their self-punishment on its head and motivate themselves to exercise with the thought (or something like it): 'I'll feel sore after exercise, and I want that'.

3. Someone plans to submit a paper to a conference in the hope of having their ideas heard. When the time comes to submit, they can't bear the thought of being heard. They would much rather remain anonymous. Instead they think about a different fact that they still attach importance to: that if they succeed in this endeavour, they will feel proud of themselves for their

achievement, and for carrying through their project. Taking this as their reason, they bring themselves to submit.

It is helpful to start with feature a), since it is weaker than b). It is clear, I think, that in each of these examples the agent is deliberating whether or not to perform a certain action. In example 1 the agent is deliberating whether or not to go through with baking an apology cake, and they find a reason that motivates them to go through with it. In example 2 the agent is deliberating whether or not to exercise, and their reflection on the desired state of bodily soreness is decisive in that deliberation. The agent of example 3 is deliberating whether or not to submit at the crucial moment. In their deliberation they reflect on the pride they would feel on submitting and they decide to submit on this basis.

If b) is true, that means that each of the agents in these examples is deliberating *because* they intend to do the thing which they are deliberating whether or not to do. There are of course different ways of filling out the examples, but on some of them this would be true. Take the agent of example 1. They are too grumpy to make reparations; they can't be bothered. One of the things they *could* do is straightaway abandon their plans to bake the cake. But instead, we can imagine that they regard this, with a twinge of disappointment, as a bad outcome: they would prefer that they bake the cake, because at some level they do want to make reparations. So instead of simply abandoning their plans, they deliberate on whether or not to go through with it. They can't force themselves to go through with it, so they have to find reasons that will motivate them. This they do through deliberating further on the merits of baking the apology cake. In deliberating, they act in accordance with their prior plans to bake the cake, plans to which they are emotionally attached. They deliberate on whether or not to bake the cake, hoping to find motivating reasons, because baking the apology cake figures in their plans: because, that is, they intend to bake the cake. Though they are inconveniently grumpy, if they can help it, they will. Fortunately, they succeed in finding this new, motivating reason.

Similarly with the agent of example 2. They planned to exercise. Now they can't get themselves to do it. Must they abandon their plans? Not obviously. Perhaps they are very experienced in this exact situation and have faced it many times before; neurotics often face repetition of this kind. So we can imagine that they know what to do, and they are confident that they can get themselves to go and exercise by thinking through further whether to exercise or not, because they are confident that by thinking about it they will find *something*, some reason, that will motivate them to go. Ergo, they do still intend to exercise, and they aim to use their deliberative powers to fulfil their intention. So they deliberate whether or not to exercise; and they deliberate because they intend to exercise, despite their recalcitrant motivation.

Example 3 is a little different. Examples 1 and 2 featured reasons and motives that were not very good reasons and motives to act on: spite, and a desire to punish oneself. But the opportunity to feel pride *is* a good reason to act. So we can imagine someone intending to submit a paper, but finding that they just don't seem to be able to click the 'send' button. Realizing this, they step back and think further on what reasons there might be to submit, hoping to use their deliberation to overcome their freeze. So by deliberating whether or not to do it, they realize that they will feel pride if they do, and this motivates them. Their motive for deliberating was so that, by deliberating, they could perform the action that their deliberation concerned whether or not to do. They deliberate because they intend to submit.

I stress that these examples are certainly not simple *illustrations* of action as a conclusion of deliberation (if they were, then that thesis could be defended through example alone). If they are examples of action as a conclusion of deliberation, then they are examples of an intention's containing a presumptive answer to the deliberative practical question, where the action constitutes the actual answer; they are not examples of reasoning just before the action itself. The point, rather, is that they cast doubt on proposition 1 of our inconsistent trio, and so help refute what I take to be a good argument that casts doubt on the possibility that action is, categorically, the conclusion of some deliberation even where a future-directed intention holds. Their relation to that thesis is indirect.

This solution makes some presuppositions: most notably, it presupposes that, in order for an agent to rationally intend, they need to have sorted out the rationality of doing what it would be an intention to do. Notably, the reason that motivates their intention in each of the examples is not a state-given reason but an object-related reason, one that bears on the worthwhileness of doing what would be intended. The agent of example 2 considered exercising worth doing, and therefore worth deliberating about so that they would do it. Exercising figured in their plans, with deliberation as an instrumental means to achieving it. And if something figures in an agent's plans, they are intending to do it. So a basic rendering of the practical reasoning involved is available: exercising is worth doing, therefore it is worth including in my plans, so I hereby plan to do it. Still, the deliberative question the formation of an intention answers is 'what shall I intend/plan?'.

Of course, this proposal is actually a basic consequence of our version of the doctrine that action answers the practical question 'how to act?' when it concludes deliberation. That requires that that question is left open by prior intention. Hence, the adoption of prior intention cannot be grounded in the closing of that very question. This just requires supplying an alternative account of the rational basis of intention than that it answers the question that action otherwise answers. If it is a cost, then it is just a cost of the thesis that action concludes some deliberations.

It may be objected that, in each of the examples, the agents aren't truly deliberating; they're doing something that merely closely resembles deliberation, such as trying to find additional reasons to act than the ones they already know, airing their motivations for acting otherwise, and so on. The objector here would press: they are not deliberating whether to do something on the basis of intending it, but rather deliberating so as to shore up a crumbling intention or crumbling motivation.

However: if an agent is trying to find reasons to φ, and working through reasons not to φ, and they know that if they find reasons to φ, they will φ in normal conditions, and they know that if they fail to find reasons to φ, they will fail to (intentionally) φ, then there does not seem to be a principled basis

on which to deny this activity the status of deliberation. The fact that it occurs within and because of a plan does not mean that it is not deliberation.

An agent can think about what reasons they would have to φ in a hypothetical situation, without thereby being in any way motivated. Likewise in a non-hypothetical situation they may be able to think about further reasons against φ-ing, as an intellectual exercise, even where they have already firmly decided to φ. But where their φ-ing and their not φ-ing depend on whether, in this thinking activity, they find enough reason to φ or not to φ, then this activity is deliberation.

The objector who accepts most of the above but doesn't call it "deliberation" seems to have reduced the issue to a verbal dispute. The substantive claim is that: *if* whether or not an agent φs depends on whether they find reason to φ that they can then act on, and they have not yet found that reason, *then* there is an open practical question for them, whether or not to φ.

Our construal of our practical cases would be inadmissible for doxastic deliberation. If you believe that *p*, then there is no open question left; no way for you to deliberate whether *p* is true or whether to believe *p*. Your believing *p* involves treating those questions as answered. The closest you can get is deliberating whether there is better evidence for *p* than you currently take there to be. There is no analogue in doxastic deliberation of needing motivating reason to φ in addition to having good reason to φ. There is no motivating evidence as distinct from evidence that you recognize as good evidence. This is part of what makes this objection difficult to parse fully.

It might also be objected that the whole point of the practical reasoning that leads to the formation of an intention is to sort out one's attitudes to the reasons to φ. So only an irrational agent could intend to φ and then perform more practical reasoning on that very subject, which is what the agents in the examples do.

There is indeed something irrational in the neurotic agents in my examples. Though they take themselves to have reason to do what they are intending to do, they just can't bring themselves to do it without thinking up further reasons: their responsiveness to reason is limited. Perhaps that suffices for an

ascription of irrationality. But irrationality in *that* sense, of limited responsiveness to reason, is something we probably all have to live with. It would be astonishing, I think, if our ideas of the practical dispositions involved in intention were benchmarked to an agent who utterly lacked that kind of irrationality. Habitually akratic agents have just as much need of intentions, and remain equally capable of action. Intentions ought to retain their practicality even in these cases. I have suggested a way in which they might do so. For the agent who handles akrasia competently, the formation of an intention sets them on a course which makes them likely to perform some preferred action, even if at the time of the formation of the intention, they cannot articulate to themselves reasons which presumptively settle the question. Their relationship to the reasons for which they will perform the future desired action is, at the time of the formation of the intention, opaque. They don't necessarily know what they are. Yet the intention may be productive in finding them. We should prefer a concept of intention that is tailored to accommodate some kinds of irrationality.

So we should deny that this counts as an objection even if the constitutive thoughts are true. In any case, there is some reason to think that they are false. Bratman has argued that the ideal of self-governance provides a reason for an agent to continue to accept their prior plans and policies in further practical deliberation[114] and to treat them as binding in certain ways. He may be wrong about the significance or nature of this reason, but it is hard to deny that there are some reasons relevantly similar to this. If an agent has intended for a long time to, for example, take revenge on an old enemy, then it is something of a shame if, having formed and executed a perfect prior plan, they back out at the last minute. There are many versions of this kind of story in which they have excellent reason to back out at the last minute: perhaps they realize the costs their revenge would have on innocents, or something like that. But it is hard to deny that the sheer dominance of their prior intention in their activity up until this moment, its role as a guiding project in their life, has at least some weight. We can call this kind of reason an 'augmentative reason': an augmentative reason to φ pertains only because, independently of its

[114] Cf. Bratman 2014; Bratman 2009

obtaining, there is reason to φ, or an intention to φ that is responsive to the independent case for or against φ-ing.

Case 3 can be read as amenable to similar justificatory structures. That the agent has intended to submit a paper provides some reason for doing it, because not to do it, having previously intended to do it, would be to back out, and that is worse than considering the whole issue on the spot and deciding not to submit anything. Intention, then, does provide some extra reason to act *in special cases*, namely where the intention has been significant in the life of the agent up until the time of action. The point is that this kind of reason to act is something that would not pertain in the absence of a prior intention, and that suffices to secure the general result that, having formed a prior intention, there is sometimes more practical reasoning to be done.

The existence of this kind of reason connects with the justification for the thesis that action is the conclusion of some deliberation, outlined earlier. Since this reason requires the existence of a prior intention, any reasoning conducted on its basis could not conclude in an intention, on pain of that reason being radically idle. Since it is a reason for action, then an agent can act on it, and can come to act on it through practical reasoning performed after the formation of the intention.

Conclusion

This chapter has argued that there are independent reasons to think of intention as failing to answer (in a sense anything more than presumptive) the question 'What to do?'; this helps to motivate the idea that it is a true answer to the question 'What to intend?'. If this is right, then it follows that it is, at most, only part of the agent's response to their reasons for action – but directly responsive to their reasons for intention. This brings into play important questions about the relationship between reasons for action and reasons for intention that are addressed in the following chapters. For it seems that reasons for action can be among the reasons for intention, yet not in such a way that 'What to do?' collapses into 'What to intend?'. What is required is

an account that provides for a conceptually distinct yet intimate relationship between these two sets of reasons.

5. Intending and Trying: towards a theory of intention

Intention can be considered in the light of two of its constitutive connections: as the successor to desire, or as the precursor to action. That it should exist at all shows something important about the nature of rationality in action: it seems that if rational or reasons-responsive desire alone were sufficient for rational or reasons-responsive action, then intention would be at most a relatively superfluous element of agency. Conversely, the importance of intention either shows something of the inadequate powers of desire, or else it shows that the tasks of rational agency were more numerous and heterogeneous than we thought they were, so that even rational desire cannot satisfy them all[115].

Plausible and attractive though this argument is, this chapter argues that it is mistaken. This idea of intention as the mediator between desire and action, it turns out, contains a tension, at least when that idea is interpreted in a certain way. This is traceable not to any distinctive feature of intention but is rather an aspect of the *place* of intention among the psychological antecedents of action as occurring after all-things-considered desire and before action. Consequently, a parallel tension pertains also to the case of trying, which, likewise, occurs (if it occurs) after all-things-considered desire and mediates between that desire and action. The central suggestion made here is that the debate concerning intention and the debate concerning trying can be read synoptically, and that this comparison is centrally useful for thinking about intention.

The Davidsonian tradition in the philosophy of action stresses verdictive desire as the central fixed point among the psychological antecedents of action, desire helping to constitute action through causing it[116]. This chapter

[115] The former alternative is taken by Pink 2016, ch. 10 ; the latter by Bratman 1987, ch.1 . Pink's view implies a rejection of all-things-considered desire in favour of a conception of *intention* as what responds to 'the full range of justifications' (p.188), and consequently falls outside the range of views centrally examined in this chapter.

[116] Cf. Davidson 2001, p.47-8.

too takes all-things-considered desire for granted as a viable starting-point (while remaining neutral on the other elements of Davidsonian action theory): questions of desire's relation to reasons-judgments, as well as the distinction between all-things-considered and all-out desire, are not addressed here. Instead, this chapter focuses on the relationship between all-things-considered desire, intention, trying, and intentional action.

This chapter has the following structure. First, I describe three plausible ideas in the theory of intention and show how a certain interpretation of them brings these ideas into contradiction. A coherent position on intention must then deny the conjunction of these ideas; I show how this structural background issue generates problem cases for various contemporary theories of intention and describe how these theories have responded, or could respond, to this. I then make the case that a similar tension pertains to the theory of trying and show how a popular contemporary theory of trying deals effectively with the issue by adopting an interpretation of one of the ideas that dissolves the contradiction. I then sketch the parallel theory for intention, and end by describing two areas where further research may illustrate the extent of this theory's usefulness.

Important general ideas in the theory of the nature of intention

Not all of the attractive ideas that generate the tension are concrete claims: what is perhaps the key proposition expresses rather a principle of research. Hence there is a tension insofar as there is an intelligible tendency for research that accepts all three ideas to run into certain problems whose nature is illustrated in advance by how these ideas, abstractly formulated, interact. Regardless of whether the claims involved are accepted by the relevant theorists (I suggest later that the principal motivation for rejecting each of them is precisely the inconsistency with the others, interpreted a certain way), they form a useful framework in identifying some of the costs that I shall argue are active in various theories of intention. They are claims one would expect a positive reason for denying.

1. The pervasiveness in normal circumstances *of intention. For the most part, wherever an agent intentionally φs, they intend to φ.*

A simpler version of (1) asserts that intention is *ubiquitous* with respect to intentional agency: that an agent intends to φ if they intentionally φ. There are a variety of putative counterexamples to this ubiquity claim, and it is important accordingly that an account of intention not imply full ubiquity. Nonetheless, the weaker claim made here that intention is *necessarily normally* involved in intentional agency is widely accepted and thus forms a more appropriate basis for identification of a motivating tension in the theory of intention. This section discusses the counterexamples to the stronger ubiquity claim in order to give some indication of what sort of restriction is implied by 'normally' as it occurs within this weaker, pervasiveness claim. The purpose of this section is not to evaluate the counterexamples but to explain the meaning of and motivation for (1).

The counterexamples purport to show intentional agency without (the right sort of) intention, and the most important class of these relates to expressive action[117], whose key feature is that they are not done under the auspices of a desirability characterization of the actions themselves: the fact that the agent 'felt like doing it' exhausts their own conception of the basis of their action.

Certain rational requirements appropriate to full intention are intuitively inapplicable to expressive actions. Suppose an agent kicks a stone out of frustration and inadvertently smashes a neighbour's window: they can be criticized for acting irresponsibly but not (except as a joke) for having failed to aim more carefully. This latter class of criticism would relate, if valid, to the agent's failure to select a plan of action preferable to the one they had, one that would have taken into account the likely effects of careless aiming. The inapplicability of that criticism reflects the fact that the kind of action the agent performed, the expressive kicking of the stone, was not internally subject to rational planning requirements. At most they should have been externally subject to the influence of the agent's other plans through active policies of self-restraint. This indirect connection ensures that criticism of the

[117] Cf. Hursthouse 1991; Chan 1995

agent is still possible but that it does not take the form of plan critique. But the agent still intentionally kicked the stone, despite the absence on their part of a plan for doing so.

The connection exploited here between intention and planning emerges also in a second class of counterexamples that suggest that side effects of intended actions are sometimes intentional without being intended[118]. For example, if an aircraft bomber pilot intends to bomb a munitions factory, knowing that doing so will inevitably kill fifty civilians as a side effect, then if they go ahead with the plan they kill the civilians intentionally – or so this style of counterexample goes. Irrespective of the issue of the validity of such examples, the important point is the parallel with the case of expressive action: because the side effect here is not what the agent plans or aims to achieve, it does not count as intended, but the claim made is that intentionality is consistent with this.

These two classes of counterexamples differ in that the expressive action case posits intentional action without any relevantly related intention present in the case, whereas the side-effect case posits an intentional φ-ing without an intention to φ. This feature is also present in a final putative counterexample to the simple ubiquity claim, Bratman's video-game example[119], in which an agent plays two games at once in an attempt to win each but, owing to a peculiar set-up of the machines, is able to beat at most one. Bratman claims that if they beat either game then they do so intentionally, but that they cannot rationally intend or plan to beat each game. Without rehearsing in detail the reasoning for the example, it is important to note that this is not like the side-effect case in that the intentionality of beating either game is not derivative on an intention to do something else of which it would be an effect. Instead, it is product of factors that do not constitute intention: 'I want to hit [the in-game target] and so am trying to hit it. My attempt is guided by my perception of the target. I hit the target in the way I was trying, and in a way that depends on my relevant skills. And it is my perception that I have hit it that terminates

[118] Cf. Knobe 2003; Harman 2006 for discussion.
[119] Cf. Bratman 1984.

my attempt.'[120]. Nor is it a case of expressive action: it is a distinct class of counterexample.

Given that in each example the absence of agent's purpose or plan was an aspect of the claim that there is no intention[121], or no intention with the right content, a useful working hypothesis is the following common diagnosis: that intentionality occurs without some intention when the agent acts intentionally despite having no corresponding purpose in their action[122] (taking purpose to imply the presence of a plan). This explains why intention should be necessarily normally present when an agent acts intentionally, for it is necessarily normal that when agents act they have some purpose in acting; and it explains why the divergence in these cases between intentionality and intention should be restricted to special circumstances rather than reflective of some deeper contingency of connection.

2. The substantiveness *of intention. If an all-things-considered desire occurs, it occurs without an intention thereby occurring.*

In order to accept (2) one must accept the non-identity of intention with all-things-considered desire; and one must also accept a commitment is to the possibility of divergence, so that it is possible for an agent to have an all-things-considered desire and not to intend. A theory of intention must explain why (2) is true, if it accepts it.

The substantiveness of intention is easily linked to a general desire to be anti-reductive, and this aspect has been much emphasized by Bratman, who aims to combine an anti-reductive conception of intention with a broadened understanding of the needs of agency in conditions of cognitive limitation, so that intentions can be understood as helping to optimize agency to function in

---

[120] Ibid., p.381.

[121] At first sight this seems an odd description of the videogame case, in which it is taken that the agent is genuinely trying to win each game. But trying does not imply purpose: it cannot be *simpliciter* the agent's purpose to beat each game. If it were, the agent would be criticizable, for their purpose to beat game A is inadequately served if they are, at the same time, trying to do something that, if successful, would frustrate that purpose, namely beating game B, and *vice versa*. Rational requirements appropriate to purpose therefore support the general diagnosis.

[122] This of course leaves opaque the nature of intentionality itself in relation to purpose; though see O'Shaughnessy 1980b, esp. ch.10, 67-8, for a suggestion.

such conditions. Instead of being merely the products of desire, intentions are 'conduct-controlling pro-attitudes'[123] that are constitutively related to the agent's plans[124]. Plans are themselves 'typically partial'[125], in usually only committing the agent to some general course of action without filling in all the details of enactment. The capacity to make and make use of partial plans represents a distinctive mode of agency, dubbed 'planning agency'[126] which Bratman claims is 'a general mode of functioning…for which there are powerful reasons'[127].

The general desire to be anti-reductive about intention expresses a principle of research more than a philosophical claim about intention which we might be in a position to definitively accept or reject. This idea of intention as *sui generis* explains why (2) is true, but is also not required for (2). I do not offer (2) here as an indubitable claim about intention but rather as an attractive principle which, as I shall show, contributes towards a central theoretical tension. But this anti-reductivism may not be universal[128] and it is important to note that endorsement of (2) often rests on more directly substantive claims concerning intention itself.

An example is Setiya, whose core concern remains with the explanation of the nature of intentional action. He is motivated by the thought that all-things-considered desire and its precursors are *not* sufficient to explain all those aspects of intentional action that need explanation[129]: this obliges the postulation of another, distinctive, intermediate mental state, not reducible to the former, a state which constitutes intention.

Although Setiya's concern is with the doxastic relation of an agent to their action and the capacity of intention to explain this feature of action, one may

---

[123] Bratman 1987, p.16.
[124] Ibid., 29.
[125] Ibid., 29.
[126] E.g. in Bratman 2009a
[127] Ibid., 232.
[128] Davidson's work is an exception in its reductivism: cf. 'Actions, Reasons, and Causes', in Davidson 2001.
[129] In his case, the claim that 'when someone is acting intentionally, there must be something he is doing intentionally, not merely trying to do, in the belief that he is doing it'. Cf. Setiya 2007, p.26.

also more directly claim that intention itself involves some unique belief that is not contained in all-things-considered desire, for example, the belief that one will φ that many suppose to be an element in intending to φ[130]. Accepting that this or some other similar belief is an element in intention automatically commits one to (2). (2) is endorsed particularly clearly by analyses of intention as a kind of belief or as an amalgam of a belief with something else.

Both of these motivations reflect a common commitment: that the notion of all-things-considered desire cannot capture what is distinctive about intention. Recognizing this common commitment allows us to gloss over the various means by which it is enacted in favour of getting clear about the more general theoretical tension.

3. The sufficiency *of all-things-considered desire. Once an agent possesses an all-things-considered desire, it is possible for an intentional action to ensue from that desire if circumstances do not prevent the possibility of the agent's acting.*

One way of understanding the connection stated in (3) is through notions of rationalization or reasons. If an agent has an all-things-considered desires to φ and they take themselves to be justified in that desire, then it seems that they *must*, rationally, take themselves to be justified in φ-ing: the justification-connection between these two things cannot diverge. To put it another way, an all-things-considered desire to φ is justified just so long as there is no other course of action ψ-ing that one ought to do over φ-ing; similarly, φ-ing itself is justified just so long as the agent does not take themselves to have reason to do anything else over it. Being justified in the same way, the justification of the one entails the justification of the other.

Once this is accepted, we get almost to the proposition represented in (3): for assuming that agents can (unless circumstances inhibit them) do what they take themselves to be justified in doing, it follows from the above that if they have an all-things-considered desire which they take to be justified then they can do what it is a desire to do. But (3) does not state that in order for an all-

[130] E.g. Velleman 1985.

things-considered desire to render intentional action possible the agent needs to consider that desire justified: accordingly this explanation of (3) does not quite capture it in its full scope.

A similar point applies also to a more minimal conception of action on which action is only intelligible if it can be given a desirability characterization (where this may conceivably fall short of what the agent believes is a reason for them to perform that action)[131]. Since all-things-considered desire too must carry with it a desirability characterization of the thing it is a desire to do, we have again an unbreakable connection between all-things-considered desire and intentional action: the existence of all-things-considered desire closes the gap between an agent's being able to φ and their being able to φ intentionally. Whether this fully captures (3) turns on whether one assumes that desires must carry with them a desirability characterization[132]: rejecting this idea would leave open the question whether if an agent has an all-things-considered desire then intentional action may ensue even if it lacks a desirability characterization.

Though the proper scope of (3) may in this way be subject to dispute, the core possibility claim in some form remains common ground. Nothing in the following arguments shall turn on whether an akratically formed all-things-considered desire itself brings in train the possibility of intentional action on its basis, circumstances permitting; the puzzle I shall raise goes through even if we restrict the proper application of (3) to non-akratic all-things-considered desire. So we shall proceed taking (3) as true enough for the purposes of argument.

The tension here

These three propositions are not inconsistent. By (3), once all-things-considered desire is reached, it is possible for intentional action to ensue thereby. By (1), if an intentional action ensues it normally comes with an

[131] Cf. Anscombe 2001, section 37.

[132] E.g. as suggested in Setiya 2007, p.62.

intention. It follows that once an agent possesses an all-things-considered desire, it is normally possible for them to form an intention to act (subsequently I will take the 'normally' qualification as read). This is in no way inconsistent with (2), the idea that such an intention is formed separately from the all-things-considered desire that would precede it. This conjunction of ideas, far from being inconsistent, expresses rather a core commitment of much practical philosophy: the idea that once an agent forms an (let us suppose, justified) all-things-considered desire, it is rational for them then to intend that action, attempt it, and if they can, perform it; and that if they are rational, these are all things they can do. This set of ideas is positively attractive.

The substantiveness of intention is a general claim that might be unpacked in a variety of ways. A definite implication is that the conditions required for the existence of an intention are not present in virtue of the conditions required for the existence of an all-things-considered desire and its precursors. But there is a separate idea that is *not* an implication of the foregoing, but that is closely related to it: that the conditions required for the *formation* of an intention are not guaranteed to obtain by the conditions required for the existence of an all-things-considered desire. What this slightly subtle difference amounts to can be expressed in the following way: the latter idea entails the possibility of cases in which an agent possesses an all-things-considered desire but is unable to form a corresponding intention. But that is not an entailment of the former idea, the definite implication, which suggests only the possibility of cases in which an agent possesses an all-things-considered desire but does not yet have a corresponding intention. The former idea is perfectly consistent with the claim that, if an agent possesses an all-things-considered desire, then, necessarily, conditions apply that make it possible for them to form the corresponding intention.

Any way of unpacking (2) must therefore explain how it is possible for an agent to possess an all-things-considered desire but lack an intention; but (2) may also be unpacked in such a way as to imply that it is possible for an agent to possess an all-things-considered desire and yet be unable to form an intention. When this holds, there is then an inconsistency. For (1) says that

agents cannot intentionally φ unless they intend to φ: so if an agent cannot intend to φ, they cannot intentionally φ. So an agent who has an all-things-considered desire to φ but who cannot intend to φ cannot intentionally φ from that desire. This runs counter to (3), which asserts that having an all-things-considered desire to φ makes it possible for you to intentionally φ from that desire if you can φ at all.

A tension holds between the three propositions, therefore, as long as there is a temptation to unpack (2) in a way that implies that there are cases in which an agent possesses an all-things-considered desire and can't form the corresponding intention. And there is a temptation of exactly this sort, as I shall now show; and these cases are, as we would expect, well-known problem cases for the views involved.

What makes the following cases problem cases is that, according to the views of intention involved, agents are prevented from forming intentions that, by (1), would be needed in order for them to act intentionally, even though they have an all-things-considered desire to perform the relevant actions and can perform them. Thus, their inability to form intentions narrows the field of their agency. In what follows, it is rational agents who are uniquely prevented from forming intentions in this way. The issue in these cases is that, owing to the conception of intention involved, there is no way in the circumstances that the formation of the relevant intention would be rational.

For each of the conceptions of intention involved, we can construct cases in which an agent possesses an all-things-considered desire to φ but cannot rationally form an intention to φ. Such cases violate (3): for it seems that regardless of what else is going on, if an agent all-things-considered desires to φ then they may, without hindrance to their rationality, intentionally φ.

This general dynamic applies to all views of intention on which a belief that one will φ is an element in the analysis of intention (hereinafter referred to as *cognitivist* accounts). On this view an agent cannot rationally form an intention to φ when they cannot rationally believe that they will φ. There are well-known problem cases for such accounts.

For example, suppose an agent all-things-considered desires to win a sprinting race, but doesn't believe that they will: they're quite pessimistic about their relative speed. But they enter the race anyway. As it happens, they try their hardest and win: it seems that their winning was something they intentionally did. But this occurred without the agent believing that it would, and accordingly, a cognitivist cannot account for the agent's intention to win. Davidson's carbon-copy case[133] is, temporal differences notwithstanding, structurally similar.

Davidson intended and treated this case as a counterexample to cognitivist accounts: the agent intends, but without believing that they will do it. However, there is a another, less intuition-reliant, point that they can be used to make: if the cognitivist is right against their opponents in thinking that agents cannot (rationally) have intentions in cases of these kinds, and if the cognitivist also accepts the ubiquity of intention to intentional, purposive action, it follows that the agents in these cases cannot act intentionally (if the actions are purposive, which, in these cases, they are). This implication is more firmly implausible than the suggestion that agents in these cases do not have intentions. It points to a deeper puzzle concerning agency: how it could be that an irrational agent would be more able to do what they have reason to do than a rational agent, if a rational agent cannot intentionally do what an irrational agent may be able to intentionally do.

Unless the cognitivist rejects the claim that intentional, purposive action requires intention, their only option in dealing with these cases is to make a belief that one will act rationally accessible wherever one has an all-things-considered desire[134]. If the viability of this position can be established, then this will be a genuine way to reconcile (1), (2) and (3). But the very mundane case of not expecting oneself to be able to do something, and then doing it

[133] Cf. Davidson, 'Intending', p.92, in Davidson 2001.

[134] This reply is more plausible in respect of cases of intention in spite of consistent past akrasia. It is suggested in Moran 2001, ch.3 section 2, that agents must see their actions as being up to them: seeing them as possible things they could do, it is open to them to take responsibility for doing them, and so acquire in this way a belief that they will that reflects only this fundamental presupposition of agency that the actions are up to them.

intentionally and on purpose anyway, suggests rather strongly that the cognitivist will instead have to drop a commitment to (1).

What is important here is a more general analytical point concerning the structure of the analysis of intention: it is *inevitable* that these problem cases would exist for these cognitivist views given the root inconsistency between (1), their interpretation of (2), and (3), all of which have some basis in intuition. For to return to a point made above, all-things-considered desire just is sufficient to render intelligible and intentional action done on its account: so to require more for intelligible intentional action than what all-things-considered desire provides is to impose standards that some cases may conceivably fail to meet. What is 'extra' in intention on these accounts is, like a belief that one will act, just contingently related to the all-things-considered desirability of an action. Such accounts can therefore be counter-exampled, by examples of intelligible intentional action that feature all-things-considered desire but that do not feature the additional elements that intention is supposed to involve: the ensuing debates merely trace the lines of this inconsistency.

Unsurprisingly, parallel problems emerge for those aspects of intention that are additional to all-things-considered desire apart from a belief in the performance of the action. Bratman claims that intention is a mental state formed with an eye to controlling future deliberation, in particular, preventing reconsideration of the relevant issues. Intention goes beyond all-things-considered desire precisely in having this deliberation-controlling effect, which constitutes it as a 'commitment'[135] over and above desire.

Bratman accordingly faces a difficulty explaining the relevance of intention to spontaneous action[136], where no issues of future control arise. The difficulty, to put it in our terms, is that a rational agent does not have any need of intention in the spontaneous case: how it is rational to form an intention in that case is not clear. Yet if spontaneous intentional action requires an intention then this means that it is not clear how a rational agent could perform

[135] Bratman 1987, p.4.
[136] Ibid., p.126.

a spontaneous intentional action, even if they have an all-things-considered desire to perform that action. And this runs contrary to (3).

The best reply available to Bratman would presumably be that forming intentions in accordance with one's conception of one's reasons to act is, irrespective of the contingencies of deliberation control, a reasonable habit of intention formation[137] and so agents do behave rationally in forming intentions just where they have good reason to act accordingly: intention for spontaneous action that reason supports is then possible. But equally, a more fine-grained habit of forming intentions for the more restricted range of cases in which agents have need of intention's power of deliberation control would also be a reasonable habit of intention formation (if such control really is intention's fundamental usefulness). Thus having a policy of not forming intentions in the spontaneous case is also rationally permissible: so it may be that a rational agent cannot, because they have that policy, form an intention to act now even where they have an all-things-considered desire to act now. Thus the connection between all-things-considered desire and corresponding action would be weakened undesirably by the agent's rational policy of intention formation[138].

The problems don't end with specifically spontaneous action. Suppose it happens to be true of one's situation that one has decisive reason to φ in the future and yet one knows that nothing in the way of control over future deliberation is needed to secure one's φ-ing because one's desire will get one there of its own accord. An agent is thirsty in a desert: they see an oasis: they intensely desire to go there and drink. By Bratman's criterion, since the agent does not need to control their future deliberation in order to ensure that they will drink, the intention to φ is merely rationally permissible to adopt (and the corresponding policy of intention formation merely rationally permissible to have), so that the agent will not break their conformity to reason by having it. But it seems, more strongly, that an intention to go there and drink would be positively justified: it is an intention the agent has good, perhaps even

[137] Ibid., p.52.

[138] It seems instead that having decisive reason to φ ought to be enough to justify an intention to φ: a suggestion developed by Shah 2008.

decisive, reason to adopt (as with the corresponding policy of intention formation). How a rational agent gets from all-things-considered desire to intention in such cases is not clear if Bratman's criterion for the rational formation of intention is encoded as a restriction on the rational formation of intention within a corresponding policy of intention formation, which, as argued, must be rational on Bratman's conception of intention's point.

To reiterate the general diagnosis: if (2) is interpreted in a way that suggests that there is more to intention than there is to all-things-considered desire and that this extra component consists in some mental state whose conditions of rational formation are not necessarily co-extensive with the conditions of rational formation of all-things-considered desire, then it follows that sometimes the agent may be in possession of an all-things-considered desire and yet be unable rationally to form an intention. If acting intentionally requires the formation of an intention then a rational agent will be unable to act intentionally in those cases, contrary to (3) and the motivations lying behind (3).

Some further theoretical options

So: as we have seen, once (2) is unpacked in this way, an inconsistency emerges. And since this sort of way of unpacking (2) is standard in the literature, it therefore becomes possible to analyse the literature in terms of the stand they take in dealing with this inconsistency. Since they cannot coherently endorse all three of the propositions – (1), modified (2), or (3) – instead they must deny at least one. So we can confirm our interpretation of the theoretical choice point involved by analysing some of the positions taken, and disagreements pursued, as if they were motivated by a need to respond to the inconsistency we have outlined.

Davidson solved the problem by rejecting (2) outright. Davidson's conception of intention endorses pervasiveness and the sufficiency-of-verdictive-desire thesis at the cost of the substantiveness of intention. Davidson suggests that the ascription of the relevant beliefs and desires that explain action also amounts to an ascription of intention: that intention is all-

out desire (in turn, it is with reference to all-out desire rather than all-things-considered desire that Davidson would understand (3) to be true). In turn, these are jointly explained as constituted by a subset of 'the same genus of pro attitudes expressed by value judgements', namely, those that 'are distinguished by their all-out or unconditional form.'[139]

Cognitivists about intention may deny that their interpretation of (2) leads them into the inconsistency I have identified: it is open to them to argue in the way sketched earlier, so as to suggest that all-things-considered desire of its own accord licenses the extra, cognitive element in intention. Velleman, at least, seems to entertain a different position, one on which intention and the capacity to act intentionally involve much more in the way of extra elements. Velleman suggests that acting on an intention involves acting on a desire to act in an intelligible way – namely the way in which one's other desires point[140]. *This* desire, it seems, cannot be guaranteed by the mere existence of all-things-considered desire itself: so Velleman already denies (3), by weakening the connection between an all-things-considered desire to φ and φ-ing from that desire.

One option for a Bratman-esque view is to deny or weaken (3) – but, as discussed above, that would be to inhibit spontaneous action as well as other kinds of action; surely an unattractive outcome. But if Bratman were to allow the possibility of intentional spontaneous action, and of intentional action in the other circumstances we canvassed, while holding onto his account of the substance of intention, he ought to abandon the claim that all intentional action is accompanied by intention. This repair is suggested by Herdova[141], who argues that since spontaneous actions do not require anything in the way of the kind of weighty, heavy-duty intention that is Bratman's paradigm, there is no reason to believe that such actions are performed out of an intention. (Herdova herself puts her claim more aggressively, suggesting that since the Bratmanian conception of intention is correct, spontaneous actions aren't performed out of proximal intentions). Since (1) is stated relatively modestly

[139] Davidson, "Intending", in Davidson 2001, p.102.
[140] Velleman 1985, p.41.
[141] Herdova 2018

in terms of the *pervasiveness* of intention rather than its ubiquity, this may not seem much of a cost: it simply represents a further, weakened, version of (1).

To make this move would be to endorse (2) and (3) and to save the theory by rejecting (1). The materials for such a move exist already within Bratman's work in his distinction between planning agency and non-planning agency[142]. (All we need in this context is the idea that planning agency is a way of going on that involves acting according to intentions and the plans they reflect, whereas non-planning agency is a way of going on that involves only acting on the basis of beliefs and desires that don't have anything to do with plans). The ubiquity of intention for intentional action would hold true of an agent who conformed perfectly to the norms of planning agency, always integrating their all-things-considered desires into partial plans and then adjusting relations between the new and the old partial plans accordingly. Likewise, it would never be true of that agent that they acted straight from an all-things-considered desire without ever forming an intention in between. However, if human agents are sometimes non-planning agents, then that means that they sometimes intentionally act despite failing to integrate that action into their planning; and on the tight link between intentions and plans on which Bratman relies, that should lead to a denial of the ubiquity of intention. To read the problem cases in this way is to suggest that rational agents are able to act intentionally in such cases by retreating from the resources, and the demands, of planning agency, and that this is something that the peculiarities of the cases permit them to do rationally.

For the causal theory of agency, on which actions are such in virtue of being caused in the right way by antecedent mental states, and in particular for any version of that theory that claims that intentional actions are such in virtue of being caused by intentions, these problems are fundamental: unless an adequate account of intention is given, the account cannot say what actions are. The pervasiveness of intention emerges as a core commitment and this puts (2) and (3) into contention.

---

[142] Bratman 2009a, p.227-8.

I emphasize that all of the problems I have discussed stem from a common source: the acceptance of (1), (2) and (3) along with a particular interpretation of (2). If the objections I have identified cannot be got around, then the inconsistency requires us to abandon one of the above. The obvious option is to try to interpret (2) in a nonstandard way, a way that does not commit us to accepting that sometimes an agent possesses an all-things-considered desire and can't form the corresponding intention. But the appeal of postulating extra belief-like components in intention makes this difficult to carry off.

However, the tension I have identified in the theory of intention is not uniquely a tension in the theory of intention. Below I shall argue that it is also a tension in the theory of trying, and that it likewise lends itself to the same derivative inconsistency as pertains to intention. But there the way to avoid the inconsistency is more obvious; and on the presumption that general problems merit general solutions, this will give us the essential clue to resolving the problem in the case of intention.

An analogous problem for trying

Like intention, trying is a member of the chain that comprises the apparent antecedents of action. Trying, if it occurs at all, comes after intention and prior in the causal-explanatory order to the action itself: this is its characteristic place. And as with intention, the relation between trying, its antecedents, and the possibility of action can be called into question.

1`. The pervasiveness in normal circumstances *of trying*. *For the most part, whenever an agent acts intentionally, their action is one they have tried to perform.*

There is a simpler version of this claim that has been defended[143]: that trying is ubiquitous with respect to intentional action, so that an agent has tried to φ if they intentionally φ. This ubiquity claim for trying is more popular than the corresponding claim for intention, and would also suffice for the following argument, but it is important to register a kind of direct counterexample that

[143] Hornsby 1980, ch. 2.1 (p.34) provides a classic defence,

is relatively convincing, along with a more general source of skepticism about either the ubiquity or pervasiveness claim.

Schroeder is the most well-known exponent of the kind of skepticism I have in mind[144]. He suggests that whether it makes sense to talk of an agent's trying depends on the speaker's perspective and is not fully determined by what pertains to the agent's own side. Agreeing roughly with the view of Hornsby (1980) that it would normally be appropriate to describe an agent as having tried to perform some action only when 'for some reason or other, the agent did not – or it was thought that he did not – straightforwardly and easily φ'[145], he takes this (*contra* Hornsby) to capture this condition on conversational context.

If Schroeder is correct, then the analogy between intention and trying that I am arguing for is inapt: with intention, but not with tryings, a pervasiveness claim would hold that is an ingredient in a relevant theoretical tension. His claim will be unconvincing to anyone who believes that tryings are particular actions[146]. The satisfaction of Hornsby's condition cannot bring into existence an action that wasn't there before; it follows that sometimes agents try to perform actions even when it is not (or not yet) felicitous to talk of their trying (and even when they are not intentionally trying to do something). The fact that we don't usually consider it appropriate to talk of agents as having tried to do what they have simply and successfully done will not be a powerful argument against the pervasiveness or ubiquity of trying once this is granted.

There are also intuitive cases of successfully and intentionally doing something 'without even trying': having a good time at a party, effortlessly skating across an ice rink, or relaxing into oneself while meditating. One is tempted to draw a distinction between 'trying' in the sense of putting in an effort and 'trying' in some other, more philosophically relevant, sense. But the possibility of such cases is some reason to reject the ubiquity claim in

[144] Schroeder 2001

[145] Hornsby 1980, p.34.

[146] Which is not everyone who accepts pervasiveness/ubiquity; Ruben is an example of someone who denies that tryings are particulars but accepts ubiquity nonetheless. Cf. Hillel-Ruben 2016, p.272.

favour of a pervasiveness claim, and I will remain neutral on the genuineness of such cases. Later (in footnote 149) I offer a suggestion as to what such cases gesture at.

2`. The substantiveness *of trying. If all-things-considered desiring and intending occur, they occur without an agent thereby trying to act in the way they intend to*.

(2`) is guaranteed to be true as long as future-directed intention and future-directed desire are included under the scope of what is mentioned in the antecedent of this claim. 'Trying' starts only when the action itself would start and this means that a gap must hold in principle between the onset of future-directed intention and the onset of trying. (2`) I shall take as indisputable when it is read in the intended sense.

3`. The sufficiency *of prior elements in the chain. Once an agent possesses an intention, it is possible for an intentional action to ensue from that intention if circumstances do not prevent the possibility of the agent's acting.*

(3`) expresses a rejection of the claim that only a class of phenomena more restricted than intentions are capable of leading to the occurrence of corresponding intentional action. The connection here seems fundamental even if spelling it out fully requires recourse to notions of reason and their relation to an agent's motivation.

Again, these propositions are not inconsistent. Perhaps when an agent has an intention, they are thereby in a position to come to perform the intended action (by (3`)). This requires them, by (1`), to try to perform that action; and it does not follow that they must already be trying to act in virtue of having the intention, leaving this consistent with (2`). And this set of ideas, that trying occurs after intending in the rational-explanatory order and itself explains action, is just what is involved in the idea of the chain of the psychological antecedents of action.

However, (2`) can be interpreted in a way that does entail an inconsistency. (2`) on its own implies only that it is possible for an agent to intend to φ while

not trying to φ. This should not be confused with a separate claim, that it is possible for an agent to intend to φ and yet be unable to try to φ.

What account of trying would exemplify this claim? Suppose one thought that trying was equivalent in some way to the activation of the will to act, where mere intention is not so equivalent. Then one could conceive of an agent's being somehow debilitated in the operation of their will in spite of their intention to act, so that when the time comes to act they cannot activate their will and so they cannot act. Ruben suggests that 'it can be that someone has a final stage intention, has the ability and the opportunity to act on it, there being no preventers or blockers, but does not do so.'[147] In denying that 'memory loss, weakness of the will, or forgetfulness' count as preventers or blockers and so fail to fall under the category of inhibiting circumstances to action's possibility, Ruben implicitly accepts the falsity of (3`). If weakness of will doesn't count as a preventing circumstance, then there will be cases in which weakness of will is so severe, entrenched and reliable that it is just not possible for an agent to act, despite having an intention and despite the absence of preventing circumstances.

However, an alternative extant conception of the nature of trying enables us to sidestep the tension. If we want there to be some substantive aspect of trying, some conditions involved in its existence that go above and beyond the conditions involved in the existence of intention, then there is a way to guarantee that it doesn't interfere with the power of intention to make action possible: and this is to say that whatever is 'extra' in trying is present in action itself. If we say this, then to say that intention has the power to make action possible is also to say that, insofar as it makes action possible, it makes trying possible too: and so it is possible to try only if it is possible to act. This rules out the possibility of cases in which an agent cannot act, despite having an intention, because they are independently prevented from trying. Instead, cases in which an agent cannot act because they cannot try are cases in which the intention fails to make the action possible: they are cases in which circumstances prevent trying because circumstances prevent acting.

[147] Ibid., 285.

Therefore, we can consistently accept (1`), (2`) and (3`) while denying that it is possible than an agent should intend to φ and would be able to φ were it not for some independent impediment to trying to φ.

This solution depends on unpacking (2`) in terms of a constitutive connection between trying and action: one that implies that intending to φ *does* guarantee the possibility of trying to φ if φ-ing itself is possible. This connection is implied by a prominent position in the literature on trying: that trying to φ is *doing what one can to φ* (for that reason, that it is what one can do to φ)[148]. Nothing more can be involved in doing what one can to φ than what would be involved in φ-ing itself: even where one tries to φ and fails, what one intentionally does in the course of trying is still part of what one would do in φ-ing[149]. So whatever the nature of the connection between intending and trying, it cannot impute more in the way of potential barriers to trying out of an intention to act than what we might already suppose to pertain to acting out of an intention to act. Taking (3`) to capture the latter connection, therefore, it follows that there are no extra conditions required for an agent to try to act when they intend to act than are included in (3`). Hence intending to act may lead to trying to act when there are no preventers or blockers to the agent's acting.

This offers some support against Ruben's claim that to count memory loss, weakness of the will, or forgetfulness as blockers to acting from an intention would 'stretch the concept beyond recognition'[150]. Those circumstances are just circumstances which would internally inhibit action itself: an agent who

---

[148] Cf. Hornsby 1995, p.530; O'Shaughnessy 1973, p.369.

[149] This account validates the pervasiveness of trying to intentional action. It is necessarily normal that when one intentionally φs on purpose one does what one can to φ (for that reason, that it is one what can do to φ). It may seem that it validates, even more strongly, the ubiquity of trying to intentional action, for it is difficult to see how one can intentionally φ on purpose but not do what one can to φ (for that reason, that it is one what can do to φ). The no-effort cases mentioned earlier may still be accommodated, however, since this analysis of trying leaves open the possibility that though an agent φs on purpose, not everything they do to φ is done for that reason. Perhaps the automaticity or unthinkingness of the action disengages it from reason (and it is just these cases that are plausibly instances of no-effort cases). Thus, though the agent purposively φs, and they do what they can to φ, not all of what they do to φ is done for that reason, that it is what they can do to φ. It may be this feature that explains the inapplicability of notions of trying in such cases.

[150] Hillel-Ruben 2016, p.285.

had started acting would be unable to continue if those things occurred[151]. They are blockers to action. Since they *directly* prevent action itself, they derivatively prevent trying to act and so prevent trying out of an intention to act.

There is a difficulty with this general conception: the well-known case of an agent with (unknown to them) a paralyzed arm, who cannot lift it but who does respond to an order to raise their arm by trying to raise it[152]. They try to raise it and nothing happens. Hence, they try, but apparently without performing any action. This is a putative counterexample to the account.

This conception of trying attractively allows us to sidestep the tension posed by (1`), (2`) and (3`). Its key characteristic is the interpretation of the substantiveness of trying in terms that make it clear that it cannot restrict the possibility of actions. It achieved this by explaining the nature of trying derivatively on an understanding of actions and what they comprise: whatever it takes to φ is what, in virtue of doing some portion of that for that very reason, allows an agent to count as trying to φ[153]. Taking intentional, purposive agency as prior to trying in this way is what allows the problem to be neutralized.

But general problems should have general solutions: and that which allows a non-threatening interpretation of (2`) might also be presumed to allow a non-threatening interpretation of (2), the analogue for intending. Since the

---

[151] This speaks to Ruben's intuitive explication of what a 'blocker' is: 'A blocker is like a blockage in a water pipe. The water flows to some point in the pipe but no further, because there is a blockage…the preventer or blocker has to interrupt a process that would otherwise have run to completion.'

[152] The case is an old one; but see Hornsby 1995, p.531, for a recent discussion.

[153] Doing what one can to φ for the reason that it is one what can do to φ does not entail that the agent's purpose, in doing what they can to φ, is to φ. The purpose may alternatively be to see whether one can φ, or to show that one can't. For example, someone who exists in the world of Arthurian legend may know that they are not the chosen King and may be motivated to prove this to everyone else. So they try to pull the sword from the stone. Their purpose is to publicly fail, thus proving that they are not the chosen King: but this strategy only goes through if what they are trying to do is to pull the sword from the stone: to do the very thing which, if they could do it, would prove them to be the chosen King. A structurally similar point, concerning the mismatch between trying and ultimate purpose, can be made with respect to Bratman's videogame case (cf. above).

essential puzzle with trying and intending is the same, we might expect that a solution in the one case should be roughly mirrored in the other.

Constructing a similar solution for the theory of intention

What allowed a solution in the case of tryings was not merely the convergence of conditions that would prevent trying, on the one hand, and those that would prevent action, on the other, given a suitable intention. A general convergence of that sort does not require (3`), the thesis that intention makes action possible if there are no preventing conditions to action. Committing ourselves to (3`), it was requisite to understand the convergence of conditions inhibiting trying with those inhibiting acting in terms that implied that there could be no barrier to trying that would not also be in a more direct way a barrier to acting: in effect, in terms that made the latter set of conditions explanatorily prior. This section explores what it would mean to offer a similar account for intention, one that would reconcile (1), (2), and (3). Other than the interest of such an account in resolving that tension, it is of secondary interest in the way it would preserve the parallel with trying. As I have insisted, the essential problems with intending and trying are similar: it is a test of this thesis that for every attractive position in one debate, a parallel attractive position can be constructed for the other.

If to try to φ is to do what one can to φ, then trying is an action. Intention is not an action: there is nothing one does in φ-ing such that, in doing that thing, one counts as performing an act of intending. Nonetheless, the priority of intentional, purposive agency that played such a crucial role in the solution for trying is useable here. I suggest that intention is the state one occupies when some relevant part of what one does intentionally is done for the sake of that which is intended[154]. This means that when an agent intends to φ, what they intentionally do out of that intention is equivalently what they intentionally do for the sake of φ-ing. Hence, the formation of intention just

[154] In what follows, I restrict myself to intentions to act. But I take it the basic idea could be applied to intentions-that: they too involve the agent in doing (or where applicable, not doing) what is needed to ensure that the intended state of affairs comes about.

marks the point where activity begins, or at least may begin, that admits this sort of teleological explanation.

This constitutes an interpretation of (2): being in a state such that teleological explanations are applicable to some of what one does is to go beyond what is involved simply in the possession of all-things-considered desire. All-things-considered desire does not itself involve its agent in acting on a purpose; but once the agent has an intention, then they cannot fail to have a purpose.

In relying and taking for granted notions of purpose, and in particular of the purposes of agents in their actions, this interpretation of intention may seem strongly to disappoint the ambitions of the causal theory of agency. But although the teleology of action is taken in this account to be explanatorily prior to intention, the teleology of action may still be given an interpretation within the causal theory of agency in terms that do not circularly invoke intention. An account is still available, for instance, purely in terms of belief and desire (taking these to be distinct from intention as per (2)), so that an intention to φ would be then construed as the state that obtains when an agent acts out of the relevant belief and desire motivating them to φ by intentionally doing things that they take to be means to or ways of φ-ing. Equally that account is avoidable: the suggestion regarding intention I have made here is neutral on the fundamental nature of intentional, purposive agency. My claim is that the capacity to act for the sake of an end is primary (relative to intention) and its exercise, when the end is itself an action, constitutes an intention to perform that action.

Intention is efficacious: sometimes agents perform actions because of their intention to perform them, so that the intention plays some role in the explanation of how the action came about[155]. This account suggests that the

[155] If Bratman is correct that intentions (like, or as a species of, commitments) involve some inertia, so that agents with intentions are less disposed to reconsider, then this very feature of intention may sometimes explain why actions happen: sometimes an agent acts partly because they didn't reconsider their action and this in turn because that that action was something they intended. We should see this as an aspect of what agents are capable of: intentions have inertia and are accordingly efficacious insofar as agents with intentions themselves purposefully don't reconsider those intentions and so end up acting on them. This purpose-involving analysis of inertia may either be taken to illuminate or obscure the phenomenon; this is an issue for further work.

efficacy of an intention to φ is constituted by the agent's own efficacy in their capacity to come to φ through what they can do for the sake of that end. This is a consequence of the constitution of intention by intentional, purposive agency. In exactly the same way, the efficacy of trying in securing the successful doing of what one is trying to do is just to be understood as the agent's power to do that thing through whatever it is that constitutes their trying.

Thus, the extra element in intention that is absent from all-things-considered desire is necessarily something present in intentional purposive agency itself. So for all-things-considered desire to render intention possible is also for it to render action possible. The interposition of intention, then, between all-things-considered desire and action cannot result in any further possible impediment to acting on all-things-considered desire than the relationship between desire and action considered independently of intention.

Because our understanding of intention draws on an understanding of intentional, purposive agency, our account of the conditions that we take to block the formation of intention given all-things-considered desire is derivative on our account of the conditions that we take to block the occurrence of action from all-things-considered desire, as (3) states. If all-things-considered desire does render action possible subject to the absence of circumstances inhibiting action, it follows that it can never happen that an agent should possess an all-things-considered desire but be unable to form an intention to φ (unless such conditions obtain).

In the final section of the chapter I wish to clarify the implications of this view of intention for two issues: the ubiquity assumption captured by (1), and the possibility of pure intending.

Further consequences

(1) stated that intentional action necessarily normally involves an intention to perform that action. I acknowledged and accepted various counterexamples to a simpler ubiquity claim but retained the pervasiveness claim as an

assumption on the basis that none of the kinds of counterexamples to the ubiquity claim would offer a way out of the puzzle cases inevitably generated by the criticized interpretation of (2). The question remains whether the account of intention I have offered validates the right version of the pervasiveness claim without falling foul of the counterexamples to ubiquity. Were it to be subject to similar counterexamples, that would be an independent objection against the account irrespective of whether it neutralizes the tension between (1), (2) and (3).

Fortunately, the suggestion I have offered regarding intention's nature appears to validate the pervasiveness claim offered as a substitute for ubiquity, on which intentional action occurs without intention when it occurs without purpose and with intention otherwise. If intention is a state roughly constituted by the agent's having a purpose and intentionally doing things for the sake of achieving that purpose, then the account predicts that intention necessarily exists only for agency that is both intentional and purposive. Accordingly, what we should expect if this account is true is that the counterexamples to ubiquity are remedied when we include a purposiveness condition: this is the diagnosis of (1) argued for earlier. Thus, a strength of the account is that it validates the correct version of the pervasiveness claim as well as giving a deeper explanation of its truth.

Let us now turn to the issue of pure intending: intending that occurs 'without practical reasoning, action, or consequence'[156], so that, *inter alia*, the intention exists without any action yet undertaken out of that intention or with that intention. I have defined intention as a state in which relevant kinds of teleological explanation are available for the relevant part of what an agent does. But the agent in a state of pure intending is not yet doing anything out of that intention: consequently there is nothing in the way of action to explain. The applicability of the above definition is now rendered unclear: in what sense are teleological explanations 'available' if there is nothing there for them to explain?

[156] Davidson, "Intending", p.83, in Davidson 2001.

Structurally, however, our ambition to give parallel accounts of intention and trying is aided rather than obstructed by this objection: for a similar objection is available also with respect to tryings, in the form of the paralyzed-arm case mentioned above. The issue with the paralyzed-arm case was that there was nothing that the agent *does* other than 'try to move their arm' as directed: consequently this was an objection to the account of trying as doing what one can. An agent who does not do anything does not do what they can in the course of trying: nonetheless, they try, hence an apparent inadequacy in the definition. Such cases exemplify a 'pure trying' that is exactly analogous to pure intending.

The task of reconciling this account of intention with the possibility of pure intending is taken up in the next chapter (the comparison with trying is left behind). It is argued that the essential point is to follow through on the idea of intention as an exercise of agency that occurs not necessarily in the context of any action, but rather consists in a state that makes teleological explanations applicable. It will be seen that this idea can be made substantive without reference to the idea of a mental action intrinsic to the adoption of intentions. This chapter has prepared the ground for a theory of intention that detaches it from the necessary roles usually ascribed to it, of applying judgments of action to action, or as aiming at good action.

6. The overall view of intention as an exercise of agency

The overall purpose of this part (chapters 4-6) of the thesis has been to rebut at a deeper theoretical level the claim that it is constitutive of intention that it is answerable to judgments about the worthwhileness of what it is an intention to do – the view that, roughly, if it is worthwhile to do that thing, the intention is then to be counted as correct or well-formed, and ill-formed or misguided otherwise. This position implies that if an agent rationally forms an intention to φ then they must be sure enough that φ-ing *is* worthwhile. Our overall view of intention requires this to be false; hence the variety of arguments supplied in the previous chapters against this position. Along the way various claims have been made about what intention is and isn't. The purpose of this chapter is to review and synthesize these claims, and to indicate what sort of constraints they place on a positive theory of intention, and what sort of theory of intention could satisfy them.

A start was made on that task in the last chapter. There I argued that we can get quite far in a theory of intention just by holding onto a few basic ideas: the pervasiveness of intention within intentional action; the substantiveness of intention relative to all-things-considered desire; and the idea that any action that is rationally all-things-considered desirable must be rationally doable and intendable too. I argued that this led to a picture in which the substantiveness of intention relative to all-things-considered desire could consist only in something intrinsic to intentional action itself – such as the exercise of agency involved in intentional action. So, I suggested, intention is to be understood as the state one occupies when one's agency is disposed in the manner characteristic of intentional action. Specifically, one's agency is disposed teleologically, so that if one intends something, then explanations of what one does from that intention take the form: done for the sake of φ-ing – for that which is intended.

This assertion that there is just such a psychological state of this sort – a state such that, when one occupies it, teleological explanations of one's activity make sense – itself requires some explanation, however. I noted in the

previous chapter that because of the possibility of pure intending, no definite action or event need occur for an intention to exist; so if intention makes teleological explanations make sense then that must *not* imply that if an intention exists then there is some definite action or event available for them to help explain. The notion of an explanation's 'making sense' or 'being applicable' must not be allowed to have this implication.

Given the possibility of pure intending, and the identification of intention with a state that makes teleological explanations make sense, everything therefore hinges on the idea that teleological explanations of this sort *would* work for certain actions *if* they happened. This counterfactual is the only viable interpretation of that idea. What it means is this: if, for example, an agent intends to paint their room, then, if the agent goes out and buys a can of paint, then the following is true: the explanation of that paint-can-buying as done so that they may paint their room with it is an explanation that the existence of the intention guarantees *could* be true – assuming that it was *this* intended act of painting rather than another that it was done for, and so on. That is, the truth of that explanation requires that intention to exist. It would not be enough that the agent desired to paint their room: that would not suffice for the explanation's applicability.

Though their intention enables such explanations, the agent need not actually buy a can of paint or do anything else in order for us to be able to make sense of their intention to paint their room, or to acknowledge the existence of that intention; but the *distinctive* thing about this very intention to paint their room is that it makes sense of any paint-can-buying that does happen.

This conception of intention says what intention is in a very indirect way – through suggestions as to the intelligibility of certain action-explanations in situations that are not necessarily actual. This indirectness seems suspect. If intention has the power to make intelligible whatever activities occur under its guidance, and if that is its distinctive feature (even if no relevant activities actually occur), then it seems that it must consist in something substantive, something that helps to illuminate the nature of this rationalization. But what could this be?

To put this another way: above I said that an agent has an intention when their agency is appropriately disposed – but what does 'disposed' mean in this context? Since no action need occur for an agent to have an intention, this disposition of agency cannot be understood as action itself – but apparently a second, more mysterious thing running alongside an agent's action that is supposed to be the basis of an understanding of intention. This does not seem to be much of an understanding.

My overall account here is weakened by the fact that other conceptions of action and intention do seem to offer a more concrete explanation. In chapter 4 (on action as the conclusion of deliberation) I argued against conceptions of intention as a particular sort of propositional attitude (directed on questions of the form "What to do?") on the grounds that no such conception could capture the idea of action, rather than intention, as a categorical conclusion of deliberation. Yet such conceptions do make clearer the capacity of intention to make teleological explanations of actions make sense. If intention is or contains an attitude to a proposition regarding action, then that potentially offers a way to understand the teleological structure of action – by relocating it inside intention, and in particular inside the content of those action-directed proposition(s) that intention constitutively involve(s).

Davidson's conception of intention[157] seemed designed to capture this exact point by locating a means-end relation in the content of the belief-desire pair which is supposed to be responsible for the action. So, to continue with our earlier example, the agent's purchase of a can of paint was done with a certain intention in that it is due to the agent's desire to paint their room and their belief that buying the paint can will enable them to do that. The judgments that intention involves, on this conception, may themselves represent means-end structure as applying to the intended actions. (More strictly, on Davidson's considered account[158], the intention to buy the paint just is the judgment that buying the paint is all-out desirable; but this too, of course,

[157] Davidson, 'Intending', p.86-7; also, 'Actions, Reasons and Causes', both in Davidson 2001.
[158] Ibid., p.96 onwards.

contains an implicit reference to the means-ends relation on account of which this is supposed to be true).

Once intention itself is thought to contain teleological structure, it is easy to capture the idea that one's actions admit teleological explanations. The fact that those actions are motivated by intentions just so structured is enough to link those actions to an underlying teleology. And this captures also a sense in which a teleological explanations might be 'applicable' even if no action actually happens, as with pure intending. Teleological explanations are applicable just in the sense that an intention exists that contains in its content the right teleology and which can motivate actions. When that's true, it's true that that sort of explanation *would* work for certain actions *if* they happened – our earlier interpretation of what we considered the key explanandum. This apparently counterfactual property has now been understood in terms of an underlying structural mental arrangement – and that is a significant strength of views like Davidson's.

It is not all propositional-attitude conceptions of intention that are rejected here – just a notable subset. Intention, for all I have said, may well involve *some* propositional attitudes constitutively. It may, for instance, involve a judgment on the question 'What to intend?'. Also, as argued in chapter 4, it may involve psychological attitudes that are less than judgment, such as a presumption on how one will answer certain questions through one's action (which is not itself a sophisticated question, rather one resembling 'what to do?'). Rejecting all conceptions of intention as attitudes towards specifically action-involving propositions means *inter alia* rejecting the possibility of accounts along Davidson's, or other, lines. This then leaves unexplained the teleological structure of action, since it does not explain it terms of a teleological structure inhering in the intentions that bring about that action. This now seems to be a significant weakness of views similar to the one I am offering.

So what are the options for addressing this lacuna – this *nonsubstantiveness problem* that apparently characterizes our conception of intention as an exercise of agency? Teleology, as argued, would have to be central in any

substantive account along these lines. So how can the idea of intention as an exercise of agency help to understand the applicability of teleological explanations?

One possibility is that that teleological structure is just not to be explained in terms of the relationship of action to the agent's intentions. This is what we committed ourselves to in the last chapter: the idea of agency, and the associated teleological structure that is inherent in it, must be taken as basic relative to intention. This is a second reason why we are obliged to reject any account along Davidson's lines.

But even if this is accepted, we are still left with the question just raised about what the relationship of intention to action is. If intentions aren't *responsible* for actions having the teleological structure they do, then how do intentions somehow guarantee that actions that result from them can have teleological structure? How do we make more concrete the suggestion made in the last chapter that although agency is to be taken as basic relative to intention, there is still such a thing as a psychological state that necessarily applies just when such agency is in the picture?

In what follows, I fill out and make more substantive the idea of intention as a state occupied when agency is disposed in the manner characteristic of intentional action. Although the idea of being so disposed is not one I attempt to explain in terms of further ideas, it still affords distinctive explanations of certain surface features of intention, and these help to vindicate this conception of intention against the charge of nonsubstantiveness just made. The chapter does this with reference to the requirement on agents to have intentions that satisfy certain normative and predictive coherences with their other attitudes, specifically the requirement not to intend bad things and to intend good things (roughly) and the requirement not to intend things that can't or certainly won't happen. The agency-first view of intention supplies certain sorts of explanations of these coherences.

Coherence requirements for intention

In this section, I'll call 'normatively coherent' an agent who has an intention to act and whose judgments regarding how worthwhile it is to do what they intend to do are not intuitively at odds with that intention. We have already reviewed a number of conceptions of what it would take for an intention to be at odds with a judgment of that kind. In particular, we have rejected the assertion that it is part of the concept of intention that one should intend what one believes it just as worthwhile to do. On such a view, normative coherence is simple, and so I will call it the 'simple view' in what follows: it requires that the agent intend a good enough action out of those available[159]. Its conception of what intention ought to be just tracks its conception of what the agent ought to do: whatever the agent ought to do considered in its own right, so it is rational for them to intend to do. (As always, I will gloss over the problems of incomplete information and use 'rational intention' and 'appropriate intention' interchangeably in what follows).

In chapter 2 I gave one particular case (the case of the holidaymakers) which appeared to show that some actions which are on balance worse than some other available actions can nonetheless be the object of justified intentions. This complicates any attempt to define normative coherence – on the view defended in that chapter and subsequently, the appropriateness of an intention depends not only on the worth of the action intended, but also on other benefits that having that intention itself may bring. This means that there are multiple judgments with which an intention must cohere if it is to be rational: not only the agent's judgment on the worth of the action, but also their judgment on the benefits of having that intention.

The suggestion that state-given benefits matter to the rationality of an intention, however, does not mean that the worthwhileness of the intended action ceases to matter. It is clearly very important to whether an intention is rational that what the agent intends to do should be something that (they

[159] There is room for further debate here, of course, between maximizing and satisficing conceptions of what the agent is to aim at, but since I have already argued that the maximizing view needs to be rejected, given the overall argument advanced here, I won't be concerned with this debate.

reasonably believe) is appropriate enough in its own right. But the fact that state-given benefits matter to the rationality of an intention means that a more complex description must be given of what it takes for an intention to be normatively coherent in the sense defined above. This is because the idea that we may determine rational intention by *first* determining the appropriate action is dispensed with here. Yet, the suggestion here that normative coherence in this sense is a substantive constraint on rational intention implies that evaluating the worthwhileness of the actions themselves is necessary (indeed, obviously necessary) to evaluating the rationality of an intention.

Another way to frame this question is to ask how the state-given benefits of intention, and the reasons for intention thus provided, interact with the reasons to intend that are provided by what the agent will have reason to do. How may such reasons be made commensurate? How can the state-given benefits of having some intention be weighed against the benefits of acting as intended?

One possibility is *instrumentalism*[160]. On this position, the worth of the intended action matters to the appropriateness of an intention because that action is likely to be one of the consequences of that intention ('consequences' here may be conceived causally or may not; the essential point is that, on this conception, it is by intending that the agent exercises an influence on what they will do). On this view, an intention is valuable when it is likely to lead to the agent's acting in ways that are valuable, and when, balancing the likelihood and the value of this consequence against all the other consequences of the intention, the intention is therefore worthwhile – worthwhile on account of its consequence profile. Here, the worth of the intended action matters only in an indirect way to intention: indirect because it is mediated by these issues about what the intention is likely to actually cause, issues which depend on all sorts of factors that the simple view would consider irrelevant. It is important to discuss instrumentalism because, if we accept that intentions are sometimes justified by the benefits of having them, it is a natural way to make commensurate the benefits of having an intention

[160] A useful exploration of some notable instrumentalist positions is offered in Pink 1996, ch. 6.

and the worth of the intended action, both of which seem to be important in the evaluation of intentions.

Nonetheless, I take it that it is obvious to the reader that instrumentalism is unattractive. What is wrong with instrumentalism? For a start, it takes us far away from the intuitions examined in chapter one – intuitions we found initially plausible, even if argument forced us away from them. Recall the following passage:

> When we engage in practical deliberation with an aim to arriving at an intention with respect to an action, our attention *immediately* centers on the question *whether to perform that action*. There is no inferential step between the question *whether to intend to A* and *whether to A*; the former question immediately gives way to the latter. This is why we can skip the question *whether to intend to A* and start right in with the question *whether to A* and yet be recognizably deliberating about what to intend (as opposed to idly wondering whether to *A* without aiming to make up our minds).[161]

This does seem importantly right; certainly we are almost always at least not conscious of a deliberation focused on a 'whether to intend' question. With the analysis now established since chapter 1, we are in a good position to distinguish between two separate aspects of Shah's claims here. The first is an extensional claim: a claim about what all practical deliberations are and what the question is the pursuit of which constitutes that deliberation – namely, questions of the form 'whether to φ'. This was the claim that our arguments in chapter 2 strongly disputed. But second is a more important claim about the *directness* of the influence of deliberations on 'whether to φ' on our intentions. Deliberating on whether to φ and *thus* acquiring an intention if a decisive conclusion is reached – which is what Shah envisages – does seem at least what normally happens. We don't usually require or think about additional considerations having to do with the benefits, psychological, social or otherwise, of having that intention.

Yet this idea of directness goes somewhat out of focus if we accept instrumentalism – if we think that actions matter just because a correctly selected intention may bring a desired action about. Such reasonings would

[161] Shah 2008, p.5. Shah also explores what we have called instrumentalism, and that is the context in which this phenomenological point is made.

require us to evaluate the likely causal impact of our intentions, and actions only come in if we judge that they are likely to be among the effects of our intentions. But such considerations are absent from Shah's picture.

It is tempting to spell out this objection in phenomenological terms, in terms of what kinds of deliberation we self-consciously conduct and the way the idea of directness is discoverable through phenomenological inspection. But to blame instrumentalism for being at odds with what strikes us naively as being the case, to make this phenomenological objection, would be to miss a more interesting underlying critique of instrumentalism. A phenomenological objection of this kind would also be inconclusive, because it would invite the fair reply that phenomenological intuitions are unreliable indicators of psychological reality. Even if our practical deliberations do 'immediately give way' to action-centred questions, it doesn't follow that the deep structure of the agent's reasoning does that. Perhaps parts of the agent's reasoning are unconscious and not open to phenomenological inspection.

So we should leave aside that line of thought. The more interesting, non-phenomenological critique is that any deliberation along the instrumentalist lines would have us choosing our intentions while treating actions from that intention as among their likely or perhaps even their assumed effects. In being treated this way, actions from that intention would be treated no differently from other effects of having that intention – such as psychological effects or social effects. In such a deliberation one's actions are not being treated as directly in one's control, rather as something that are manipulated into existence through the selection of the correct intention. For the reasoning the instrumentalist conception has agents implicitly perform is: intention I will being about action A, action A is desirable, therefore I shall hereby adopt intention I[162].

Of course, even on this view we do in a sense choose our actions – because adopting an intention to φ is deciding i.e. choosing to φ. Yet a rational agent

[162] There is a notable similarity here to objections against volitional theories of the will (cf. Hornsby 1980, ch. 4 section 5): it is the operations of the will, here the adoption of intentions, that come to seem to be the real actions, and this itself is a reductio of the view, though one it would take far more work to make fully substantive.

chooses their action, in this sense, only insofar and in virtue of choosing their intention. They choose which intention to adopt; then, adopting it, they count as choosing an action. Yet this sort of choosing an action will not figure in their practical deliberation: that deliberation ends, on this view, with conclusions of the form: therefore I shall intend to φ, rather than: therefore I shall φ.

What instrumentalism puts out of focus is the notion that agents *in their practical deliberations* choose their action and not just their intentions. Making room for that idea means understanding the difference, on the agent's side, between their conception of their prospective actions and their conception of the other effects their intentions or actions might have: they choose the former in concluding their deliberation, yet the latter are things they may only manoeuvre into existence. Yet according to instrumentalism there is no relevant deliberative difference between the actions resulting from an intention and other effects of that intention. Thus instrumentalism must leave out a central element of the story when it comes to practical deliberation: it must be the case that agents do not see themselves as manipulating their own actions into existence, and instrumentalism offers no help with this idea. It simply leaves uninterpreted the idea of directness that is present in Shah's intuition, an intuition that we accept.

It seems that a better conception would have to do justice to the idea that it matters in a very direct way to the agent whether or not what they are deciding to do is a bad thing or not – direct and not mediated by complex judgments about what one is likely to accomplish through one's intentions. If we accept this aim, then we are committed to an understanding of normative coherence on which the appropriateness of an intention depends on something other than its likely consequences and the value of those consequences. It matters in a special way to the agent whether what they are deciding to do is a good or bad thing to do. A good account of intention would both a) explain what normative coherence is more fully, and also b) explain why agents must be normatively coherent in this way, in virtue of what sort of thing intention is.

Other than normative coherence, there is a separate rational constraint on intention: predictive coherence. We examined in chapter 3 two constraints on rational prospective intention: that is consistent with what the agent knows they can do, and that it is consistent with what the agent knows it is not the case that they certainly won't do. We argued there that the latter, which is stronger than the former, is true, so we won't revisit this issue here.

Accounting for the coherence requirements on intention: propositional attitude views

This section outlines the philosophical value in various conceptions of intention as a propositional attitude, namely, their apparent ability to explain the coherence requirements on intention. This approach is then criticized for needing to pack too much into the content of intention. This section then examines other ways of accounting for the coherence requirements without packing them into the content of intention, and concludes that though there is no decisive argument against that approach, it faces a substantial explanatory challenge.

Above we called the 'simple view' of normative coherence the view on which it is part of the concept of intention that it must cohere with the agent's judgments about what it is best to do. A notable version of this simple view is the content-based view on which it is because the intention itself (perhaps *inter alia*) a judgment towards a normative proposition that this normative coherence requirement obtains.

Predictive coherence, just like normative coherence, is something that is well-explained if the content of intention has predictive or factual implications. We can construct a common understanding of what the coherences amount to for attributions of both normative and predictive or factual content: intentions are rationally incompatible with truths of the relevant type; if intentions, in virtue of their implicitly asserted content, imply judgments of the relevant type, then they are testable against the corresponding truths; so when truths are known that contradict the content of the intention, a rational agent cannot hold onto that intention.

So, for example, we can suggest the two parallel explanations for each sort of coherence requirement: first, for normative coherence, that if an agent intends to φ and they know φ-ing is wrong, then they are irrational to intend to φ, and this is explained by the fact that their intention involves a normative judgment to the effect that φ-ing is the thing to do, a judgment which the agent knows to be false and therefore cannot rationally maintain. Second, for predictive coherence, if an agent intends to φ and they know that they won't φ, they are irrational to intend to φ, and this is explained by the fact that their intention involves a predictive or factual judgment to the effect that they will φ or at least might φ, a judgment which the agent knows to be false and therefore cannot rationally maintain.

These sorts of ideas are buttressed by suggestions to the effect that if a truth of the relevant type is asserted, then that *automatically* counts as a criticism of any intention that does not rationally cohere with it. On normative coherence, Stout writes:

> [on] the conception of practical reasoning that I am defending, [the conclusion in this case is] thinking that the thing for me to do is to eat some tripe. This can be contradicted from any reflective distance just by saying that this isn't (or wasn't) the thing for this person to be doing… we sometimes just want to challenge the conclusion of practical reasoning without criticising the process. We want to be able to say: 'I don't accept your conclusion,' without saying anything about the process that led to that conclusion… the criticiser too need not be having any thoughts about the quality of the other person's inference. I might think [that someone who intends to eat tripe] is wrong as a result of a piece of practical reasoning on my part that does not engage with any aspect of the tripe-eater's reasoning except its conclusion.[163]

If the reason why intentions have to be normatively or predictively coherent is to be located in the way normative and predictive implications pertain to stages of practical thought before an intention is adopted, then we would only be able to criticize an intention in virtue of knowing the process of its formation. What Stout points out, that such knowledge isn't necessary to criticize an intention, then supports the idea that it is something in intentions themselves that creates requirements of normative and predictive coherence

[163] Stout 2019, p.569-570

rather than anything in practical thought generally. And this makes attractive a strategy of looking to the content of an intention to explain coherence requirements.

Stout divides up possible views of intention into those that suggest that intentions are judgments with normative content (and thus testable against what the agent's normative knowledge) and those that suggest that intentions are judgments with predictive content (testable against the agent's factual knowledge). Framed in this way, Stout criticizes the latter kind of view (principally Anscombe and McDowell rather than other exponents such as Velleman) on the grounds we have seen: they cannot account for normative coherence by appealing to the nature of an intention's content, so at most they can try to account for normative coherence by making it a requirement on the process of practical reasoning – yet this is inadequate.

Yet if intention admits *both* a normative coherence requirement *and* a predictive coherence requirement, then by argumentative parity, views of the former kind – on which intentions are judgments with some kind of normative content – also invite the objection that they leave no easy way to account for predictive coherence. Judgments with normative content are not contradicted by judgments with factual content, so it is unclear how credibly pointing out that someone certainly won't do what they intend to do creates a rational requirement on them not to so intend. By Stout's own argument, his conception of intention is impugned by the existence of predictive coherence requirements on intention. And this is clearly a result of the general attempt to explain such requirements by packing in appropriate content into the intention itself, just because more and more has to be packed in: it is a problem with the propositional attitude conception of intention.

Stout has a way of making room for something that, extrapolating from his text slightly, might seem to connect with the predictive coherence requirement. It is based on the idea that judgments with the content 'I am φ-ing' appear as premises in practical reasoning that concludes in the intention to take the means to φ-ing. This structure, from a factual judgment about action to an intention, characterizes reasoning that happens 'within an

action'[164] and not just before, where action is thought of as a 'continuing process rather than as a completed event'[165]. This connects with the predictive coherence requirement because it suggests that, unless one thinks that one is φ-ing – a thought that conflicts with the judgment that it is impossible to φ and also with the thought that one is simply not going to actually φ – then a certain kind of practical reasoning cannot be successfully completed.

Interesting though this idea is, it does not quite get at the predictive coherence requirement in its connection with prospective intention. It is not just *within* the course of an action that difficulties are created when one is certain that one won't or isn't φ-ing; the difficulty comes into play before that as soon as the intention to φ is formed. It remains unclear, then, how this is supposed to ground the idea of there being a rational constraint on intention as such. To put it crudely: why should an intention be dropped simply because practical reasoning conducted on its basis cannot be completed? Why is that a problem with the *intention*, as opposed to with the reasoning or the agent or even the human condition? The extrapolation I have attempted does not yield good results: what Stout is talking about is the process of thoughtful adjustment in one's actions and this does not connect automatically to the idea of there being rational constraints on intention.

It seems, then, that any conception of intention as being a judgment with normative content is not going to easily accommodate a predictive coherence requirement. And vice versa if the content of the intention is the only way to ground such requirements. This suggests a dilemma: either intention is a factual judgment in which case normative coherence requirements are not accounted for, or it's a normative judgment in which case predictive coherence requirements are not accounted for. If this is right, then either sort of propositional attitude conception of intention is in trouble. And this would amount to a positive argument against any sort of propositional attitude conception of intention.

[164] Ibid., p.572
[165] Ibid.,p.574

Perhaps there could be a *conjunctive* account on which intentions are constituted by the combination of both sorts of judgment at once: that an intention to φ amounts to the judgment, φ-ing is desirable and I will do it, or some variation on this, or perhaps it merely involves such a judgment. That would account for the coherence requirements. But given that the relationship between each component is standardly thought to rest on practical reasoning itself (because one will φ only because it's desirable) this suggestion requires much more filling out. In any case, it is suspect to try to add in everything needing to be explained into the content of the intention; it is much more philosophically satisfying if we can explain the coherence requirements with respect to a unified concept of intention.

The general diagnosis here, that the content view, or the propositional attitude conception, needs to put too much into intention's content, assumes that Stout's argument quoted earlier is correct – the argument that suggested that the process by which intentions are formed cannot be the locus for implicit judgments that might conflict with the relevant normative or predictive truths. That is, it assumes that the only way an intention incoherent with the agent's normative or predictive judgments could be criticisable is if the intention itself involves some content inconsistent with those judgments. This argument is criticisable on its own terms even entertaining the underlying conception of intention as a propositional attitude, or as involving such attitudes. If the argument isn't right, then we can't conclude from the claim that intentions have predictive content that normative coherence requirements don't make sense, and vice versa.

Here is an alternative suggestion[166]. Intentions, on any account of the matter, are based on certain judgments of the agent's – whether or not the intentions themselves are not constituted by such judgments. This means that if such judgments are known to be false, then certain intentions must be criticisable, namely, those which are necessarily based on such judgments (or rather, necessarily based on them if they are rational[167]). In the context of predictive coherence, one could suggest that intentions are necessarily based on the

[166] It is inspired by Davidson, 'Intending', p.100, in Davidson 2001.
[167] Pears 1998, ch.9 for discussion

judgment that the agent can do what is intended or that they won't certainly not do it. If intentions are necessarily based on such judgments, then that would explain why it automatically counts as a criticism of an intention that it is inconsistent with the relevant predictions. In the context of normative coherence, one could suggest that intentions are necessarily based on the judgment that the intended action is appropriate enough in its own right. In that case (*pace* Stout) it would be clear why it automatically counts as a criticism of an intention that those normative claims are false. In both cases, an intention formed on such false judgments is thereby criticisable: the implication is that if the agent judged the matter aright, then they would intend otherwise.

Stout is therefore wrong to suggest that criticisms of the intention that don't engage with its content must implicitly attack the process of its formation or the quality of the inference; rather, they attack judgments that the agent *must have made* if they have an intention of that kind. If such a picture can be convincingly substantiated – if it can be shown that intentions must, given their nature, be based on certain kinds of judgments – then the defender of the propositional attitude conception of intention is not forced to cram into intention's content all the judgments that the agent must have made if their intention is to be rational.

One of those judgments is conceivably the idea that an action is something the agent can do. Can we devise an explanation for why intentions should necessarily be based on those judgments? It seems we can. Intentions are necessarily based on the agent's conception of what they can do because an intention is a selection of one of the agent's options for their action. And the agent's options just are what they can do.

Unfortunately, this basic story does not capture the whole picture when it comes to predictive coherence given our arguments in chapter 3. There we argued that it is not just impossibility judgments that intentions must cohere with, but also judgments that concern what the agent certainly will or won't do. These judgments, we stressed, don't impact on the agent's options – this was exactly what the Toxin Puzzle showed: the agent has the option of

drinking the toxin (otherwise there is no puzzle at all) but the difficulty for any prior intention to do so is that they almost certainly won't do it. So this easy story, just outlined, isn't quite enough to take care of predictive coherence. I cannot see any natural way of grounding the idea that intentions are necessarily based on the agent's judgments about what they will or certainly won't do. If there isn't a way of doing this, then this alternative I have been discussing fails to accommodate predictive coherence requirements, and the only option for the defender of the propositional attitude conception is to work in suitable judgments into the intention's content.

Could this approach be utilized for normative coherence requirements? This would be to say that intentions are necessarily based on certain of the agent's normative judgments. But which? Above we suggested that there are two: the usefulness of the intention, and the worth of the action. What is the story that might explain why intentions must rationally be based on these two sorts of judgments?

On the standard story we have looked at in several places, the function of intention is to guide the agent to the best of their practical options: if this is its function, then it makes sense that intentions are necessarily based on the agent's judgment as to which is the best of their options. Yet this story, as we have argued, leaves out the role of judgments as to the usefulness of the intention. On the other hand, if the agent's intentions are up to them, it makes sense that they should be selected on the basis of the agent's judgment as to which would make for the best intention; yet we argued above that judgments as to the worth of the action play a specially direct role and this is not accounted for here – so this does not help with understanding how intentions might necessarily be based on the agent's normative judgments about action in the right way, or indeed what sort of normative judgments they might necessarily be based on, given our commitment that agents may rationally intend what is not an appropriate action when just considered in its own right, or at least some actions of this sort.

Of course, there may be a more sophisticated story that the defender of the content-based approach would wish to tell about why intentions are necessarily based on certain sorts of normative or predictive judgments. I will rest here with the suggestion that the theoretical burden is now on them to show how this is true.

To reiterate our basic diagnosis of the state of play: intentions are subject to both normative coherence requirements and predictive coherence requirements. These are well accounted for if intention's content includes relevant judgments or are based on such judgments. Yet there are two basic problems: firstly, explaining what normative coherence *is* in light of our suggestion that intentions are rationally selectable on the basis of their own merits, a point that this conception offers no immediate help with; and secondly, the risk that this explanation excludes either requirement, since intention-as-a-normative-judgment would not explain predictive coherence, and intention-as-a-predictive-judgment would not explain normative coherence. Neither of these problems are decisive – but they don't need to be. They show at the least that there is no cut-and-dried story for explaining the relationship of these requirements to intention, and that philosophical progress could be made by supplying such a story.

Given the general ambition here, what would now be helpful are explanations of either coherence requirement that do not explain it in terms of consistency with content. But that requires a conception of intention as something other than or more than a propositional attitude – such as, potentially, the one we have adopted.

The key idea I shall appeal to in order to replace the work done by the notion of the intention's content is the idea of intention as an exercise of agency – an idea we first discussed in the Introduction and made problematic in the first section of this chapter.

Any exercise of agency – including action itself – can be interrupted from two separate directions. It can be interrupted from without, as when external circumstances make impossible the continuation of that exercise (given the kind of thing that exercise amounts to). And it can be interrupted or stopped

from within – as when the agent voluntarily discontinues what they are doing. These two dimensions of dependence correspond, I shall argue, to the two kinds of coherence that need to be explained. Predictive coherence is connected to the obstruction of intention through the obstruction of planning activity which that intention essentially sustains; and normative coherence is connected to the agent's refusal to will the intended action, a non-willing that itself constitutes the absence of intention.

Explaining predictive coherence

To review a point made in chapter 3 and again just above, we take predictive coherence to consist in a rational requirement on agents not to intend to do what they can be certain they won't do. The most uncontroversial instance of this is the requirement not to intend what it is impossible to do. We argued in chapter 3 that this is insufficiently broad a requirement, and that in order to make sense of the Toxin Puzzle we must suppose that the requirement extends more broadly to any action that the agent can be certain they won't perform. The real lesson of the Toxin Puzzle is that this requirement is genuinely broader than the restriction to possible actions.

The essence of the explanation of this predictive coherence requirement was already offered in chapter 3. It is essential to intention that an agent who intends will, from that intention, make (to the extent that is necessary) more detailed plans as to how to do what they intend to do. But if an action is one that they certainly won't do, then no plan will enable the agent to be confident that they will do what they intend. Hence, for any action of that sort, no plan can be satisfactory. The agent cannot plan further on the basis of the intention. The agent is necessarily wasting their time.

Here I shall expand on this sketch by showing how its key points are validated more deeply by the conception of intention on offer as an exercise of agency on a par with that involved in action itself.

The claim, one associated most strongly with Bratman, that intention is constitutively connected to plans, is something that needs independent

validation in this picture of intention. Unlike Bratman, I am not accepting the thesis that intentions *are* plans (plans conceived by Bratman as psychological states that involve 'an appropriate sort of commitment to action' as opposed to the other notion of plans as an abstract structure)[168]. If it is true that intentions just are plans, then no substantive explanation is needed as to why an intention to φ should rationally require the possibility of corresponding planning activity as to how to φ. If the agent cannot successfully come up with a plan to φ, then of course they cannot plan to φ (i.e. on Bratman's identification, intend to φ). The effect of this identification seems rather to put out of focus the need for an *explanation* of predictive coherence, since it seems to be intrinsic to the very notion of what plans are that we expect the agents who adopt them to take them to be achievable. Yet Bratman's position does successfully connect intentions with predictive coherence requirements.

Bratman's position has already been discussed and rejected in previous chapters. So we cannot avail ourselves of this sort of explanation of predictive coherence requirements. We must utilise an independent explanation. Moreover, it must be one that explains why, if no plan can be satisfactory to the agent, then the intention connected to it is necessarily *irrational*, rather than merely flawed in some other way.

When planning activity is ongoing, the realization that the action won't be accomplished is what puts an end to that planning activity. Given that we are committed to tracing such features of intention back to aspects of acting, what we need is to look for is some feature of action whereby action itself is ended upon the realization that it is not going to be carried out successfully.

This lends itself to an emphasis on know-how or skill that is exercised as one acts. To be skilled at some action or action-type is to be able to achieve it successfully in a broader range of circumstances, or to a higher standard than if one wasn't skilled. Skill and know-how are conceptually connected to success. When an action is rendered impossible, for whatever reason, the agent cannot exercise their skill successfully; the extent of the agent's skill also determines which circumstances make it impossible for them to act

[168] Cf. Bratman 1987, ch.3, p.28-9.

successfully, since a more skilled agent is able to succeed in circumstances which would make it impossible for a less skilled agent to succeed.

The notion of skill is useful here to the extent that we are prepared to accept that skill is not itself (necessarily) governed by or analysable itself back into ordinary concepts of action, so that the connection between skill and the possibility of success does not derive from the connection between intention and the possibility of success. If it did, it would not be able to sustain an explanation of the latter connection.

One thing that makes plausible the idea that it does stand independently is the point that an agent exercising a skill does not necessarily do so under the complex self-representations that we would normally associate with reasoned intention[169]. For instance, a skilled pianist is able to play a quick scale of any of the traditional forms from any starting point within that scale and from nearly any starting finger with no prior preparation, so that they need not think of what they are doing under a description more complicated than 'playing a scale of that sort'. In contrast, a less experienced pianist, before playing that scale, will have to recall the right fingering and think through in much more detail the appropriate occasions for moving one's thumb, which notes exactly are included in the scale, and so on. In each case, we can pinpoint the same action being performed, but the skilled pianist has a self-representation that is vastly reduced in complexity and size relative to that of the unskilled pianist. Though each agent is, we can suppose, successfully carrying out an intention to perform the same action, the associated intentions differ.

It is open to an opponent to suggest that the relevant intentions are merely implicit in the case of the skilled agent, whereas they must be explicitly framed by the less skilled agent, but are present nonetheless in each case. However the burden of proof is now on them to show that this is true; for the intentions that are *visible* in each case clearly differ. What is at stake is whether we can sustain a notion of skill or know-how that is not itself reducible to the carrying out of further sub-intentions[170].

---

[169] For another exposition see Tenenbaum 2007a, section 3.
[170] For further material on this topic cf. Hornsby 1980, ch. 6; Lavin 2013.

If the exercise of skill is directed at success, then action is obstructed when it is carried out in circumstances that are beyond the reach of the agent's skill. An agent who realizes that their skill is not enough here is an agent who does not know how to perform that action in their specific circumstances. Lacking know-how, they cannot perform the action (they don't know how!). This is not to imply that they need to be taught (perhaps it is impossible for anyone) but rather that the idea of their employing their skill falls flat once it becomes clear that requisite conditions are absent. Nor is to imply that their absence of particular know-how means that they lack general know-how: it is sometimes intuitive to describe an agent as knowing how to do something that they cannot do, as when someone knows how to swim such-and-such a distance but currently can't because they have suffered severe injuries. Again, the idea here is that know-how can no longer be exercised in circumstances that prevent success, and this is equivalent to saying that they lack know-how that covers these circumstances.

If this is right, then a fundamental part of the exercise of agency involved in action is the deployment of know-how in the direction of performing that action successfully[171]. And that in turn provides for a way in which to exercise agency in *that* way is to be subject to the condition that success can be expected or at least hoped for as one exercises one's agency that way. So to exercise agency in that way *before* acting is to be subject to the condition that one expects success in one's future actions through that exercise. This connects intending with planning, since it is through planning that intention can produce success in what is intended. An unsatisfactory plan is one which the agent, in exercising their agency in the direction of success in what is

[171] This general line of explanation relies on treating as central the case of *telic* actions over atelic actions: actions which are constitutively linked to some standard of success, over action which are not. Otherwise there is no ground to suppose that the exercise of agency involved in action is one that intrinsically is directed at success, rather than such direction being an incidental part of what it means to act. This is an open question in the philosophy of action; but see Lavin 2004, ch.2 for an argument.

A separate problem here is that of divine agency: perhaps God does not need to deploy know-how in order to make things happen – as Stout says, 'If it were theoretically possible to have an omnipotent agent, then this would constitute a counter-example to the claim that practical justification must involve means-end justification.' Stout 1996, p.127.

intended, will abandon in favour of a plan that does exactly that. Thus, to exercise one's agency in the same way one exercises it in acting, with intention, is to render oneself subject to the condition that it is through the exercise of that agency that success in what one intends can be hoped for through that agency – otherwise the intention is irrational. Plans are the site on which this requirement is played out, through which the agent discovers whether the planning activity that their intention motivates can come up with a plan that satisfies this condition. This is the explanation of how intentions are subject to a predictive coherence requirement.

Explaining normative coherence

As briefly described before, the theorist faces a dilemma in accounting for the requirement on agents to have intentions that cohere with their normative judgments, provided that they accept the claim I have made that the action, considered in its own right, there is most reason to perform is not always the action that there is most reason to intend.

The dilemma has two horns. For the two most natural ways of accounting for normative coherence requirements on intention are a) to suppose that the intention has a content with direct normative implications, the 'content view' presented above, or b) to suppose that reasons for intentions are explained by the likely consequences of having that intention i.e. the 'causal interpretation' presented above. Yet option a) is too restrictive: it implies the falsehood that reasons for action alone matter in determining reasons for intention. On the other hand, option b) makes reasons for action matter too indirectly: such views fail to do justice to the point that deliberation about what to do can itself often rationally result in a decision's being made without any need for an additional deliberation about what to decide to do. Both views are inadequate, and this creates a dilemma.

There is not, as far as I can see, any way to construct more ingenious versions of either view to neutralize these objections. This dilemma puts the theorist in a difficult position. The topic is the relationship between reasons for action and the rational formation of intention. How can we make room both for

sufficient indirectness in this relation to allow us to accommodate cases of rational intentions to take worse options, thus avoiding the objection to the content view, while also insisting that the quality of options itself often determines what intention it is rational to form, without any other deliberation needed, thus avoiding the objection to the causal interpretation?

The crucial point to understand is how, if intentions lack suitable normative content, their rationality can amount to something more than how rational they are in light of their likely consequences. What matters, that is, is how the rejection of the content view can leave even any other *prima facie* theoretical options other than the causal interpretation. We must understand how there can be room for something more here.

The theory of intention proposed up until this point, of intention as an exercise of agency, can help us, particularly the idea that the exercise of agency is the same in kind as that involved in action itself even if no action flows from it. From this it follows that there are certain equivalences in how the agent exercises their will with respect to both deliberation on 'What to do?' and deliberation on 'What to intend?'.

Suppose, for instance, that an agent simply cannot bear to φ: φ-ing is just too horrible and they refuse to do it. This is also to say that an agent cannot bear that exercise of agency that would constitute an attempt to φ: not only can they not successfully φ, they cannot even do what they can to φ, because if they succeed then they will have intentionally φ-ed, and this action is what they are refusing. But if this exercise of agency in respect of φ-ing is one that they refuse, then since it is the same as the exercise of agency involved in intention (*ex hypothesi*) then this refusal is also a refusal to intend to φ. Thus, a refusal to act is also a refusal to intend to so act.

This, surely, is exactly what we would expect. An agent cannot intend something that they reject as an action when considered in its own right. For them to do so would throw our attributions into confusion. Thus we can explain some refusals to intend with reference to a refusal to so act, just as we can explain some refusals to try to do something with reference to a refusal to do that thing. The most natural cases to illustrate this are cases where it is

obvious that the action itself is the source of a direct refusal, so that there is no room for the agent to modify what they are prepared to do in light of their rewards for intending to do it. For example, it makes sense that an agent who refuses to kill is an agent who will refuse to go along with plans that involve killing, and who will refuse even to try to kill.

This point is what helps us explain multiple points of interest expounded earlier. The agent of the Toxin Puzzle, we claimed, naturally refuses to intend to drink the toxin upon realizing that it will not make sense to actually drink the toxin, and while issues around predictive constraints on intention help to explain why we can consistently think of them as not being rationally required to intend to drink, there is a separate issue as to why it makes sense to think of the agent's refusal to intend to drink the toxin as being directly determined by their reflections on drinking the toxin. The equivalence pointed out above helps to explain why. Drinking the toxin itself simply has nothing going for it: there is no reason to do it and positive reason against. The agent cannot but refuse, so they also cannot but refuse to so intend.

It also connects with a key explanandum: the directness that Shah points out is characteristic of the influence of deliberation on 'What to do?' on deliberation about 'What to intend?'. When an agent refuses to do something in advance, this itself *is* a refusal to so intend: it is a refusal of that exercise of agency involved in that action and an intention to so act. Thus, on those occasions when the question 'What to do?' is decisively answered (even if only presumptively, as per our suggestion in chapter 4), the question 'What to intend?' is, correspondingly, closed. If the space of actions is so structured that only one action is one that the agent is prepared to attempt, then that also represents the sole thing that they are prepared to intend. This is why deliberation on what they are to do does have a direct influence on the agent's consideration of what they are to intend.

I have insisted that deliberating on what to do often does not settle questions about what to intend; in light of the above, this amounts to an insistence that deliberation on what to do often leaves some leeway for an agent to make a pick among some of their options without being thereby irrational. Our

understanding of what picks it makes sense to make will then hinge on a prior understanding of irrationality in action. To this point, consider the case of the incentivized child from chapter 2. The child is not that keen on studying but must form a real intention to study in order to obtain the games-playing time that they value much more. They are able to do this because they are able to rationally pick studying over some other alternative way of spending their non-gaming time. This does not require that studying is actually better than any such alternative; it only needs to be not worse enough than the other option. It can't be that the child cannot but refuse to spend their time studying; as long as that constraint is met, then the child is free to adopt an intention to study for the state-given reasons pertaining to that intention. The exact conditions of rational leeway are somewhat unclear in general terms, but there appears to be good reason to accept that there are choice situations in which some option is worse than another but where it is not irrational to pick the worse option when one has external reason to do so.

The overall definition of normative coherence cannot be stated with much precision given that the conditions of rational leeway cannot be stated precisely in general terms, but we can at least point to two key examples where the agent lacks rational leeway in what they may do: the case of drinking the toxin, where there is no reason at the time of action to drink and positive reason against, and the case of immoral or repulsive action, where the action is bad enough that an agent cannot rationally pick it no matter the benefits of intending to perform it. This suggests at a minimum that an agent has rational leeway to pick some action as long as there is a) at least some positive reason to do it and b) it is not too repugnant in its own right.

As should be obvious, this account is very much unlike the instrumentalist conception of the role of deliberation on what to do on deliberation on what to intend. It doesn't reduce the role of reasons for action in the formation of intention to what bears on the likely consequences of that intention; rather, reasons for action help to settle which intention a rational agent is willing to form in the circumstance and a selection can be made if necessary among the possible intentions that are left.

It should be obvious that this account validates Pink's notion (cf. chapter 2) that rationally taken intentions entail that the agent may carry them out without irrationality. If some action would be irrational to perform, then a rational agent would refuse it, and this would be a refusal also to intend it. What is different here is the *explanation* of the rationality-preserving aspect of decisions: here it is not explained through the conception of decision as a reason-applying activity but through an understanding of the relationship between decision and action and the equivalence of the state of the will implicit in that[172]. It is a condition of a will's being rationally employed that it is not directed at any simply irrational action, an action that the agent must refuse.

So, I have attempted an explanation of normative coherence requirements on intention that amounts to neither of the two views: the content view, on which intentions contain or are normative judgments, and the causal view, on which intentions must be evaluated in the knowledge that they will likely bring about their actions and that is the only significance actions have in the evaluation of intention. My explanation appealed to the idea of intention as an exercise of agency and absolutely requires that idea in order to work. Without it, there is no explanation of how what an agent may rationally do can have any direct bearing on what an agent may rationally intend, which we have taken as a key explanandum in the theory of intention at various points in this thesis. This also enabled us to explain the normative coherence requirement in a way consistent with the admissibility of state-given reasons for intention since we rejected the simple version on which an intention must track only the agent's preferences over their options for action. On our view, while these preferences do themselves sometimes amount to refusal to intend, they often leave room for a rational pick, and this is all the ability to adopt intentions for state-given reasons requires.

[172] Thus, I avoid being subject to Pink's charge against Gauthier (Pink makes this charge in 1996, p.176) that he has failed to explain the rationality-preserving aspect of decisions while 'arbitrarily retaining and trading on' that conception (p.173). The account I have offered provides an explanation.

Conclusion: The Stand-Taking Conception

This chapter has attempted to substantiate the idea of intention as an exercise of agency that is not a mental action, an exercise of agency in the direction of what is intended. I have done this not by trying to unpack the idea of an exercise of agency in conceptually independent terms but rather by trying to show in what directions the idea points. Key in particular was our assumption that the exercise of agency is one that is involved in action but that is not identical to action; some philosophers may reject entirely the intelligibility of this notion, but I cannot see any way to resolve a debate of that sort. It is a reflection of our agency-first view whose discussion was begun in the last chapter: the idea that we must understand intention through taking agency as something already understood.

This conception of intention I shall call the 'stand-taking conception'. Its central idea is that an agent who intends (now or in the future) to act a certain way takes a stand on what they shall do. This idea communicates our idea that, in intending, the agent is active or utilizes their powers of agency: they see it as up to them and their intention reflects a determination on this score – even if nothing actually results from that intention. Stand-taking is intrinsic to both intention and action itself and characterizes the practical.

In this chapter I have not discussed another feature of intention that is commonly thought of as something needing to be explained: the ability of intention to sustain a distinctively practical kind of knowledge[173]. I have left this alone because such knowledge pertains, in the first instance, to knowledge of what one *is* doing, rather than prospective knowledge of what one will do. To the extent that this is characteristic of action, it is characteristic of what I have specified as the exercise of agency involved in action; yet this is our explanatory concept *for* intention, not something that the concept of intention itself explains. Thus the topic of practical knowledge comes in at one stage prior to our topic of the rationality of intention adoption and maintenance, and does not need to be addressed here.

[173] Cf. Anscombe 2001; Michael Thompson 2011, in Ford, Hornsby, Stoutland (eds.); Setiya 2007, as some notable papers within a vast literature.

7. State-given reasons and instrumental reasoning: A puzzle and solution

In intending an action we aim at performing it or take a stand on our performing it, and that means that we must either have a plan for its performance, or at least anticipate forming such a plan; otherwise this creates an incoherent self-conception. I have argued that this point is paramount in forming an adequate solution to the Toxin Puzzle in the context of defending the rational admissibility of state-given reasons for intention. However, the last two chapters advanced a different set of ideas for explaining various features of intention: that intention is constituted by an exercise of agency in a specific direction. The last chapter suggested that aspects of the latter idea explain the former: why it is that an agent who has a coherent conception of themselves as intending must also have certain other beliefs and intentions. This chapter discusses these two lines of thought with reference to an objection to the intelligibility of state-given reasons for intention. It argues that, in this case, the idea of intention as an exercise of agency helps more with this problem than the idea of constraints on coherent self-conception.

The fundamental problem is that means are related to ends in a way that intentions to take means aren't related to intentions to take ends. It is through the means that the ends come about (if they do); it is by walking down the street that I get to the café, or by cycling there or getting the bus. If a means of this sort is not performed, the end is simply not achieved. It is because of this that it makes sense that there should be a fundamental principle of practical reason according to which, if the ends are justified, then there is at least one set of means for attaining those ends that is justified, and vice versa, that if there is no acceptable way to achieve some end, then achieving that end is not a reasonable thing to aim at.

This connection between means and ends grounds an intuitive principle in the theory of practical reason, one that roughly connects the justified or rational pursuit of ends with the justified or rational taking of means. There are various ways to spell out such a principle, but the fact that the taking of means is

necessary for the achievement of the end is central. Our interest is in the theory of justification within the instrumental domain. In this regard, it seems that the dependence of ends on means being taken ensures that, if there is some reason to achieve the end, then there is corresponding reason to take the means for the sake of that end. Raz has recently outlined a general principle that would plausibly govern such transmission of justification. Its central idea is that:

> when there is an undefeated reason to perform an action (the source action) there is also a reason to take any action which facilitates its performance, provided that it is part of a feasible and undefeated plan whose pursuit by the agent is likely to generate an opportunity to perform the source action, where a plan is defeated if the reason for any of its indispensable steps is defeated.[174]

This principle, importantly, connects reasons to take means with reasons to pursue ends. No part of the reason to take the means depends on the agent's actually *having* that end – rather it depends only on their having an undefeated reason for its pursuit. But its connection to the necessity of means for ends is what makes it plausible: it relies on the idea that justification transmits to that which is necessary for what is justified to be done. Raz expresses this connection when he says:

> Reasons are reasons to do what will constitute conformity with the reason… it is a reason to avoid being in a situation in which one would be in breach of that reason.[175]

A reason to perform some action, considered as a reason to avoid being in a state of affairs in which one doesn't satisfy that reason, is also therefore a reason to take sufficient means to that: to make the right preparations and perform that action. If they don't, they just won't achieve the end, and they will be in breach of that original reason[176].

---

[174] Raz, "The Myth of Instrumental Rationality", in 2011., p.148
[175] Ibid., p.151
[176] The point that the taking of means simply is required for an agent's ends to be achieved has a fundamental status in other theories of instrumental reasoning. Wallace's theory is discussed later; but also see Finlay 2014, ch.3, for an essentially predictive analysis of instrumental conditionals, on which 'ought' as it appears in such conditionals expresses probability of occurrence of the consequent on the antecedent (given the implicit background). Thus, 'in order to catch the criminal, Holmes must lay a trap' is analysed as equivalent to 'Holmes won't catch the criminal unless he lays a trap.' It is clear that the necessity of means to ends is central to this approach.

The point that the taking of appropriate means is straightforwardly necessary for ends to be achieved therefore supports useful theories of means-end justification. With it in hand, we can state a preliminary version of the problem for state-given reasons for intention: this relation between means and ends is not paralleled in the relation between an *intention* to pursue an end E and an *intention* to take the means M to it. An intention to take M for the sake of E is not required for an agent to have the intention to pursue E, in the way that the taking of M itself is absolutely required for E to occur. This is because the relation between such intentions is psychological – it depends entirely on the agent's own mind which intentions are psychologically required for them to be able to hold on to other intentions. If, for some agent, an intention to take M is not psychologically required for them to intend to E, then they may retain an intention to E without forming an intention to take M. And so any imperative on them to intend to E will not, just in virtue of that intention *itself* being required, create a secondary imperative to intend to take M. Thus, state-given justification for intentions to perform some action don't necessarily generate derivative justification for subsidiary intentions to take the means to the performance of that action. In contrast, if an intention is justified because its associated action is justified, then, because justification for actions roughly creates justification for means to the performance of those actions, intentions to take means to those actions will roughly be justified too.

The need for derivative reasons for instrumental intentions

Before interrogating this reasoning in more detail, let's first clarify why this discrepancy should matter: why it should be problematic or worrying if state-given reasons for intention should fail to generate derivative reasons for intentions that are, in their object, instrumentally related to the object of the source intention.

Consider two of the cases described in chapter 2: the case of the incentivized child and the case of the holidaymakers. In the case of the holidaymakers, the discrepancy clearly generates a problem. The holidaymakers are to intend to act so as to be ready to leave the house by 7; having a tendency to be less than

fully punctual, they are adopting this plan so that, in case they fail to meet it, they will be ready at the latest to leave by 7.30, which is the real time they need to be ready to leave by – so by intending to leave earlier, they can be sure that their mild unpunctuality won't leave them stranded. This creates state-given reason for them to intend to be ready by 7. Now for this strategy to work, once they intend to leave by 7 (knowing that they don't actually need to leave by 7), they will need to go on to adopt a wide variety of derivative intentions. For example, in order to be ready to leave by 7, they will need to set their alarms for 6 o'clock, they will need to organize the discharge of last-minute tasks and so on. So if their *intention* to leave by 7 is to have its own intended effect of galvanizing them into getting ready well in time for 7.30, that intention will have to lead to them adopting such derivative intentions. Absent the formation of such intentions, the original intention to leave by 7 won't have its intended effect of getting them ready to leave by not much later than 7, so any failure on this score introduces an insuperable problem for the justification of that very intention. The justification of that intention hinges on its capacity to play its characteristic part in producing derivative intentions to take means to the intended end.

Not all cases of state-justified intentions have this feature that their rationale itself depends on the capacity of those intentions to play their characteristic part in producing derivative intentions to take means to their associated actions. The case, also introduced in chapter 2, of the incentivized child provides a counterexample. The child must produce for their parents a credible expression of an intention to study in order to secure a dose of time with the games console. Being unable to fake such an expression convincingly enough, they have reason to form a genuine intention to study that they can then express, for the sake of obtaining said time. In this case, the intention to study only needs to be convincing – it does not directly need to be efficacious, or to generate derivative intentions regarding the studying itself. Were such intentions to be altogether off the table, an issue would arise because it would then become plausible that the child is not going to study – and so the child would not be able coherently to present themselves as aiming

at studying. Still, the connection here is more indirect than it is in the case of the holidaymakers.

If state-justified intentions to act fail to create any sort of justification for intentions to take means to the performance of those actions, this introduces a series of questions regarding how such an agent might acquire, or maintain, the requisite derivative intentions without conceiving of them as justified. These questions are, principally: how might the agent be able rationally to adopt and maintain derivative intentions to take means to the performance of some action because they intend to perform that action, where the latter intention is state-justified? Is it requisite, either for this purpose or for some other, that state-given justification for intentions generate some sort of state-given justification for derivative instrumental intentions? Are there any barriers on this score to the efficacy of state-justified intentions, of a sort that would render unintelligible those very justifications for those intentions which incorporate an expectation that, if the agent so intends, they will form and act on derivative instrumental intentions?

On some theories of instrumental reasoning, the first of these questions requires no special answer: it should make no difference whether the source intention should transmit justification to derivative instrumental intentions. Instrumental reasoning itself will still be perfectly possible and intelligible, so long as the source intention remains there to undergird such reasoning. On Broome's view, whether or not an agent has sufficient reason to intend as they do, they are rationally required, roughly, to intend what they believe to be a means to what they intend.[177]

Broome suggests that although it is possible for an agent to come to satisfy rational requirements just in virtue of the operation of 'automatic processes'[178], reasoning is sometimes needed when these are lacking: 'when automatic processes let us down, our mortal rational disposition equips us with a further, self-help mechanism…reasoning is something we do that can

[177] Broome 2013, p.159
[178] Ibid., 12.1, p.206

bring us to satisfy requirements of rationality.'[179] Instrumental reasoning is reasoning that rationality permits us to conduct in order to satisfy the fundamental rational requirement to be instrumentally coherent. Roughly, it permits us to base intentions to take means to certain ends on intentions to take those ends, provided that the achievement of what is intended is within one's power through the performance of such means[180]. No reference to reasons to act or reasons to intend is present within the conditions of the rational permission governing instrumental reasoning. Applying this account to the case of state-justified intentions, derivative instrumental intentions could be acquired through this sort of reasoning, provided that we are prepared to accept its veracity as a form of reasoning.

For our purposes, a more interesting and philosophically rich theory is propounded by Wallace[181]. Wallace is concerned in particular with the satisfaction of rational requirements in cases where the agent knows that they lack reason to intend or act as they do. He argues that agents who akratically decide to pursue some end are still capable of perfectly intelligible instrumental reasoning in pursuit of their end, despite their knowledge that they lack reason to do any of it. For example, an addict who decides to get a fix, despite knowing the damage it will wreak, may well go on to call their dealer and arrange a meeting – their instrumental reasoning here is recognizable as such. But since the agent knows that they lack good enough reason to get a fix – indeed they take themselves to have decisive reason to avoid it – we cannot construe their instrumental reasoning as, in their own eyes, coming to recognize a reason to call their dealer from their reason to get a fix.

Wallace argues that we can account for the irrationality of failing to adopt the intention to take the means to one's ends through understanding it as a form of inconsistency in one's beliefs[182]. For an agent who fails to adopt the intention to take the necessary means to their ends is genuinely landed in the

[179] Ibid., p.207
[180] Ibid, p.257
[181] Wallace 2001
[182] Ibid., p.21

following inconsistency. Intending the end, they must if they are rational take its achievement to be possible; its achievement is only possible upon completion of the necessary means; the achievement of the necessary means won't happen without an intention to take them. Without an intention to take the means, they are therefore in a position to know that the achievement of the end is impossible; but they must take the achievement of the end to be possible if they intend it.

The value of this observation for building an account of instrumental irrationality depends on the assumption that it provides a way to understand what happens when the agent is rational. Instrumental rationality, on Wallace's view, can be understood as the agent's avoidance of theoretical irrationality (which minimally involves consistency in beliefs). The agent achieves rationality by adopting those intentions that allow them to avoid inconsistent beliefs. By adopting an intention to take the means, the agent avoids having to conclude that it is impossible for them to achieve the end; they could also avoid inconsistency by dropping the intention to take the end.

If such theories are correct, then the answer to our first question – can agents acquire and maintain derivative instrumental intentions if they possess state-justified intentions to act – is, straightforwardly, yes. They do it in the normal way, moving, as is rational, from intentions to act to intentions to take means to the performance of those actions.

Examining the plausibility of these theories of practical rationality is beyond the scope of this thesis. Our question, rather, is: would the fact that instrumental reasoning is possible from state-justified intentions crowd out any need for state-justified intentions to transmit justification of some sort to derivative instrumental intentions? After all, if agents, starting with state-justified intentions, are capable of reaching derivative instrumental intentions through such reasoning, then those source intentions can play their characteristic part in the formation of derivative intentions. For intentions such as that justified in the case of the holidaymakers, there is no problem for their rationale.

There are three points that need to be made. The first is simply that there are other theories of instrumental reasoning on which it does matter that the agent should conceive of justification for the means to their ends (or intentions thereto) as being inherited from the justification that there is for those ends (or that there is for the intention to pursue the end). This chapter aims to stay neutral on which conception of instrumental reasoning is correct (including theses such as Raz' on which there is no distinctively instrumental reasoning, just a facilitative principle of reasons generation), and that purpose is defeated if the attempt is abandoned to say how justification could transmit from ends-intentions to means-intentions. This alone makes it necessary to aspire to give such an account.

The second point is that there are substantial theoretical reasons why the case of state-justified intentions should differ from the case of akratically formed intentions – reflection on which formed the motivation for Wallace's theory. It defines akratically formed intentions that they are formed in spite of the agent's conception of the overall justification they lack to so intend. The psychological force behind such intentions ensures that the agent has sufficient motivation to adopt further instrumental intentions, in order to avoid giving up their akratic intention – regardless of whether or not those derivative instrumental intentions are supported by reasons. In the akratic case, we can bracket the question of whether the agent thinks that the derivative instrumental intentions are justified: *ex hypothesi*, whether the agent thinks such questions can be answered is not necessary to explaining why they adopted such intentions. State-justified intentions, on the other hand, are notionally formed through responsiveness to reasons, so that they won't be adopted unless they are conceived as substantively justified on balance. So there is no ground here to reject the assumption that the question of whether derivative instrumental intentions are justified will matter to the agent, and will impact whether they form such intentions. If such questions do matter to an agent, then in order for derivative instrumental intentions to be formed, they must themselves be formed on the basis of being perceived to be justified. This lends support to the idea that an account of instrumental

reasoning in the case of state-justified intentions does require a principle describing transmission of justification to derivative instrumental intentions.

Matters are more complex for Broome's theory, for which the conception of such instrumental requirements and the kind of reasoning that reliably ensures their satisfaction is not grounded in reflections on akrasia[183]. Still, there is independent reason to think that some sort of transmission of justification must be presupposed. Broome proposes that agents are subject to 'the enkratic requirement' – roughly, that rationality 'requires you to intend what you believe you ought'[184]. Since what an agent ought to intend is itself a function of reasons (or the agent's beliefs about them) on his view[185], it follows that derivative instrumental intentions, if they are to be rational, cannot be contrary to the agent's reasons to so intend. So derivative instrumental intentions must be supported enough by reason, in the agent's eyes, if they are not to fall foul of this rational requirement.

This is because of the final point: that several issues concerning instrumental intentions that are derivative on state-justified source intentions require robust reference to reasons governing these intentions. When it is necessary to think

---

[183] Broome's suggestion is rather that 'often the [reasons to act] leave you free to intend something or alternatively to have the opposite intention' (Broome 2013, 5.4, p.86). Broome argues that there is no reason why the reasons to act in such cases should collectively require the agent not to possess contradictory intentions – no necessary way in which the agent, in possessing both intentions, inevitably does something they shouldn't do. To use an example similar to Broome's, suppose I intend to pack lightly for my coming holiday and also intend to pack lots of luxuries. Both of these courses of action are individually sufficiently sensible or good (let us suppose), and either is permissible and non-criticizable. If I possess both intentions I may eventually end up acting incoherently – I won't be able to have a stable policy on whether to put things in my backpack or keep them out. But the temporary possession of both intentions prior to that sort of conflict may not prompt me to do anything silly, as long as I lose them before any such conflict gets underway. On the contrary: by planning to pack luxuries, I may end up purchasing things that I will use anyway, whether or not I take them on the holiday with me, so that the net effect of having this intention alongside the intention to pack lightly is net beneficial to me. Nonetheless, as Broome suggests, having both intentions is irrational. No coherent basis is afforded for instrumental reasoning. My instrumental reasoning requires that the means I take to my ends don't exist alongside an intention to do something contrary to my ends: instrumental reasoning requires as the satisfaction of rational requirements that, if Broome's argument is correct, are more demanding than the reasons to act themselves.

[184] Ibid., 9.5, p.170

[185] Cf. Ibid., ch. 4, which argues that reasons themselves are just explanations of why one ought, or things that contribute to such explanations.

through which possible intention is justified, the resolution of this question requires thinking about which intention is the most justified. When one of the intentions is instrumentally relevant to one of the agent's background state-justified intentions, then this can only figure in the choice situation insofar as this aspect is reflected in the reasons that are taken to apply to this intention. This cannot be done unless instrumental intentions in some way inherit justification from state-justified source intentions.

So there is substantive reason to want some principle governing transmission of justification from state-justified intentions to further instrumental intentions. The next section explores a natural way of providing for such a principle.

Incoherent self-conceptions: a possible solution

One point raised in the previous section is very suggestive: Wallace's insight that an agent, unless they either have or anticipate forming a plan of action, cannot rationally believe that they will do what they intend to do. Without this, their intention cannot be rationally held regardless of how state-justified it is – conversely, retaining that intention requires forming a plan of action. This suggests (going beyond Wallace's own theory) that derivative instrumental intentions might be valuable to the agent insofar as they permit the retention of a valuable source intention. If this is correct, it means that, when a source intention is valuable, it is also valuable to the agent to form derivative instrumental intentions. So then there is exactly the principle of transmission of justification that we have been looking for.

This approach draws also on Raz' point (quoted earlier) as to what reasons are:

> Reasons are reasons to do what will constitute conformity with the reason… it is a reason to avoid being in a situation in which one would be in breach of that reason.[186]

If an agent is rational, then adopting instrumental intentions helps them conform to the reason they have to maintain the source intention. This is how

[186] Raz 2011, p.151

the reason for the source intention generates a reason to adopt further instrumental intentions.

Even from this brief rendering, it should be apparent that this reasoning would have to be essentially reflexive and psychological: it is their own intention that the agent must ensure they maintain, and they adopt further intentions just by way of anticipating being unable to so intend. Among other things, they require the concept of an intention and they require some understanding of the consequences of their own rationality. This alone makes it unlike some of the accounts of instrumental reasoning already canvassed, in which the agent is simply focused on the question of how to achieve the object of the intention. In what follows, I develop this point and suggest that this reflexive reasoning about one's own psychology, though perfectly valid as far as it goes, doesn't mirror ordinary practical reasoning and doesn't create exactly the sort of means-end justification we might expect, or want. I won't rest anything on the idea that requiring these concepts over-intellectualizes the kind of reasoning involved; I will instead suggest that the reasoning itself falls short of the mark.

We can sharpen this by constructing a paradigmatic instance of such reasoning. Of the two cases mentioned earlier, it is more helpful to use the case of the incentivized child (over the case of the holidaymakers) because as argued earlier, in the latter there is independent reason to adopt instrumental intentions but not in the former. In this case, the reasoning must be as follows:

1) If I don't now form or anticipate forming an appropriately specific enough plan for what to study, I will know that I won't do any studying.

2) If I know that I won't do any studying, I will, being rational, now cease to intend to study.

3) I must intend to study in order to obtain time with the console.

Conclusion) Therefore, I must adopt in time a plan of action for studying.

This reasoning must be reflexive because it wouldn't go through on the basis of the contents of the intention alone. For consider:

1`) If I don't pursue means M to E, I won't achieve end E.

2`) If I won't achieve end E, then I will now cease to intend E. (False)

Though 1` is true, 2` is false – we often continue to intend things that in fact we will fail to achieve – and so the conclusion cannot be detached in the way that above it was proposed that it could.

The reasoning from 1) through to Conclusion) can also be objected to on the grounds that it only goes through for agents who are already aware that attaining end E requires means M. Suppose an agent lacked this knowledge: they are in a state in which they believe that achieving E *may* require taking some means, and may not – they don't know which specific means they will be, if any, and they won't know until they research the matter. They are at present uncertain. Normal instrumental reasoning in normal cases directs such an agent to inquire into the means. If they want to attain the end, they had better be sure that they are performing the necessary means to that end: so they ought to find out what means those are, if there are any.

However, suppose an agent is merely somewhat uncertain whether attaining E requires e.g. preparations to be made far in advance. Given this uncertainty, another bit of uncertainty follows: the agent must, unless they intend to think about how to achieve E, be uncertain whether, if they fail to make preparations for E far in advance, they will know that they will fail to achieve E. In terms of the above reasoning, the analogue for this case would be as follows:

4) If I don't research how to achieve my end E, I won't know whether E requires taking a specific set of means M. (*ex hypothesi*)

5) If I don't intend M, I will know that I won't intend E.

5) is unsupported by 4). 5), which is an analogue of 1), is just not made true in these contexts of uncertainty. Uncertainty itself should not rationally break an intention; we often, quite rationally, intend to do things which we are not certain we will be able to achieve with the means we have selected[187]. This

[187] Cf. Holton 2009, ch.2

means that, in contexts of uncertainty as to which means are required, there is not necessarily any pressure on the agent who simply wishes to protect their valuable intention to achieve an end to find out which means are required for that end to be achieved. If the agent can anticipate the disappearance of uncertainty, then they can anticipate potentially being in a position where they will be unable to retain their intention with the new knowledge that they cannot achieve it. But where this uncertainty is not expected to be dispelled, the sort of reflexive reasoning we have described fails to lead the agent towards the normal practical course of researching how to attain their end.

For example, suppose an agent is financially rewarded for intending, now, to beat an opponent at chess later. They are a skilled chess player and usually win matches when they want to. This opponent could be unusual and their chances of success may benefit from, and may even require, advance researching of their typical strategies in order to be able to design countermeasures for their idiosyncratic yet predictable moves. But the reward for the intention is a reward for having the intention now – by the time the match begins, the agent will have received the reward. So if the agent neglects to research their opponent, they won't be in a position to know that they should have done until the intention to win stops being useful. This means that the agent has no reason to think that if they fail to research their opponent they will be rationally unable to intend to beat them. Thus their need to hang on to their intention to win does not require for its continuation the pursuit of such research; it is enough for the agent that they usually win. Here, the anticipation of knowledge of necessary means not taken is known not to be able to interfere with the agent's ability to hang on to their intention. But it would seem strange if the agent, intending to win, were completely indifferent to the wisdom of researching their opponent in order to increase their chances of success. This cannot reflect, then, any fear of the effect of knowing that they haven't taken necessary means, but rather something else about intending.

So this is one point at which this reflexive, psychological reasoning fails to correlate to what we would regard as ordinary, warranted practical reasoning. It reflects the fact that the implicit principle of means-end justification on

offer here only pertains to those intentions the absence of which *threatens* the source intention – it does not touch on all those intentions which, normally, it might make sense for the agent adopt if they have the source intention.

There is a second way in which this reflexive reasoning fails to issue in the full range of derivative instrumental intentions. Not all state-justified intentions are ones in which, if the agent fails to intend even the necessary means to the performance of what is intended, they will thereby be in a position to know that they won't reap the state-related benefits of having that intention. The kind of case I have in mind exploits a particular temporal structure:

Newly Opened Restaurant) A new restaurant has just opened in town. Smith derives a psychological benefit from intending to visit this new restaurant: it helps them feel like a culturally engaged citizen who is conversant with developments in their local area. This psychological benefit, let us suppose, has the following profile: it begins if Smith decides now to visit the restaurant. It persists for a month (until the restaurant stops being new and exciting) as long as Smith either a) retains, for that month, the intention to visit the restaurant, or b) visits the restaurant within that time period.

It seems clear that Smith has sufficient reason to decide, now, to visit the restaurant, but no particular reason to go soon over a bit later. He has reason to make the decision but less reason to actually visit the restaurant once that decision is taken. It is in fact indifferent whether he visits the restaurant within the next two months as opposed to later; as long as he retains the intention, he reaps the psychological benefit. He does have *some* reason to visit the restaurant, since (let us suppose) it is a perfectly good restaurant and going would be an enjoyable experience. We can suppose that he has sufficient reason to visit the restaurant in light of its qualities as a restaurant; it's not an undesirable place to dine. But this reason is disconnected from the primary reason why Smith ought to *decide* to go.

This situation is not analogous to the Toxin Puzzle; adopting the intention to visit the restaurant is not subject to the same objection that intending to drink the toxin is. In the Toxin Puzzle, it is understood that the subject will not have

sufficient reason to drink the toxin once the time comes; if anything they have decisive reason against drinking it. Here, *ex hypothesi*, they do have sufficient reason to actually visit the restaurant, not merely to intend to do so. Without some such premise, the arguments examined in chapter 3 do not go through.

Now that the basic rationality of the intention has been defended, let's turn now to the possibilities for instrumental reasoning this situation presents. In particular, let's try to see if the suggestion explored in this section can offer an adequate account of Smith's instrumental reasoning for taking the means to visit the restaurant, should he decide to visit at a particular time, say this evening. Suppose that the means to going to the restaurant involves booking a table in advance (the new restaurant is very popular). Does Smith have reason to take the means: does the means help him retain the rewarded intention, or put him in a position he needs to be in where he has that intention and is ready to do something?

Neither of these are the case. Booking the table doesn't help Smith retain the intention it is worthwhile for him to have- there is nothing stopping him retaining the intention to visit the restaurant; he doesn't need to book a table to visit tonight, or even within the next month, in order to retain that intention. So the idea that means are worthwhile because they help one have valuable intentions cannot afford to Smith a sound route for reasoning through which he can reason his way to booking a table in order to visit tonight. Yet if Smith decides to visit tonight then he ought to be able to perform instrumental reasoning as a result of which he will perform the means to doing that i.e. book a table.

So, intending to book a table isn't necessary for Smith to retain the valuable intention to visit the restaurant at some point, and by postponing visiting the restaurant, and postponing booking a table, it fails to be the case that Smith won't reap the benefits of that intention to visit the restaurant.

The issue here is that some intentions are temporally lax; they have no deadline. This means that intentions to take the means can be postponed indefinitely. Yet the state-given benefits of such intentions may be time-limited. Thus, in such cases, the formation of the intention to take the means

does nothing that could not be achieved without it. Such cases do not exemplify the means-end justification principle on offer here.

There is a final way in which this means-end justification principle fails. It relies on the assumption that one is rational: that if one fails to intend an adequate plan of action, one knows one won't achieve what one intends and so that very intention to do it will founder. If the aim is solely to retain the valuable intention, then this issue can be solved by making oneself (in this respect) irrational. That is, suppose it were possible, by the taking of an appropriate pill, to make (for example) one's intentions immune to the realization that they won't be achieved – so even if one knows for certain that they won't be achieved, one continues to so intend anyway. By doing this one would ensure that one retains the valuable intention in the face of not adopting the requisite derivative instrumental intentions. If one took such a pill, then adopting further instrumental intentions would be unnecessary. The means-end justification principle would not apply to them. Strictly the means-end justification principle would then require a disjunction in cases of state-justified intentions: either adopt necessary instrumental intentions, or else make yourself irrational. Yet ordinary practical reasoning from ends surely does not license the agent to take an irrationality pill precisely in order to avoid such reasoning.

The issue here (to re-iterate our initial diagnosis of the problem) is that the connection between a state-justified intention and further instrumental intentions is not one of necessity: sometimes agents can have the former without having the latter. Given that they are distinct psychological entities, a natural thought is that if the former is justified, then the latter is justified thereby only to the extent that it is necessary for the former to exist. Yet the version of this connection just examined is flawed. That instrumental intentions should *themselves* be conceived as means to the end of valuable intention – the means-end principle applied at a psychological level – does not ensure the full range of connections we would want. In order to ensure that state-justified intentions transmit justification to derivative instrumental intentions, we must look elsewhere.

Intention as an exercise of agency: a more helpful suggestion

In this section I attempt to apply the overall theory of intention developed in this thesis in order to solve the problem identified in this chapter. The clue was given just above: the issue is that, in our set-up of the problem, we treated the state-justified intentions as psychologically distinct from any further instrumental intentions at which the agent is supposed to arrive. Given this separation, the justification of further instrumental intentions has to be *based* on the justification of the source intention; it must inherit something of that justification, and this must be describable via some suitable principle. But we ran aground in an attempt to describe what this principle could plausibly be.

The problem disappears, on the other hand, if the underlying psychological assumption is rejected. If the possession of the source intention is somehow bound up, constitutively, with the subsequent possession of instrumental intentions when those become relevant, then the justification of the source intention automatically counts as justification of the relevant instrumental intentions. If, in other words, what is justified is some overall psychological state which includes both the source intention and subsequent derivative instrumental intentions, then there is no issue for how derivative instrumental intentions could be justified if this overall psychological state is itself justified.

The idea that source intentions and derivative instrumental intentions are constitutively linked is of course not the idea that an intention to pursue some end and an intention to take means to that end are identical intentions. They could not be identical intentions. For one thing, often there are a variety of ways to pursue a single end, and an agent may fully pursue the end in any one of those ways. An agent could have an intention to pursue an end and intend to pursue it in one particular way; they would lack the intention to pursue it any other way and this would be in no way irrational or a partial description of their psychological state.

Instead, the psychological identity has to be understood at a different level. In previous chapters, I argued for a conception of intention as a non-actional exercise of agency in the direction of what is intended. In intending, the agent

utilizes their agential powers, ones that are involved in action, and part of what defines this exercise is that powers are organized in the direction of *successful* performance. So, intention brings in its train the possession (or anticipation of such possession) of a workable plan for the achievement of what is intended. Just as the performance of the action itself is prevented by the absence of a workable method for doing it, so the intention to do it is incompatible with that too.

When there is state-given justification for an intention, this is as much as to say that there is state-given justification for the agent's taking a stand on their achieving what is intended. This exercise of agency, stand-taking, itself consists partly in the doing of what is necessary for that achievement. When an agent is so disposed, they not only intend to pursue the end; they are also disposed to adopt intentions to pursue some sufficient set of means (as far as they can see one). For intentions that are temporally lax, such as the one justified in *Newly Opened Restaurant)*, this disposition to form further instrumental intentions is itself lax; otherwise, where the end is pressing, the exercise of agency that constitutes intending will involve a disposition to form instrumental intentions forthwith.

When some exercise of agency is justified, then *whatever* the source of that justification, the agent is justified in not only intending the end, but also in adopting further instrumental intentions; such adoption is just part of what stand-taking involves. This applies to both object-given and state-given justification: if some intention is justified, so must at least some complete set of instrumental intentions be as well. This just reflects the idea that the agent, *ex hypothesi*, is justified in exercising their agency in the direction of doing something; making and following plans is part of that.

This is substantively different to the second-order approach explored above on which the need to retain a valuable intention is what drives any instrumental reasoning. If the exercise of agency involved in state-justified intentions is identical to that involved for object-justified intentions, then it will give rise to the same range of practical reasoning. Where above we objected that wanting to retain a valuable intention would not necessarily lead

one to participate in the full range of practical reasoning appropriate for intentions, here the co-extension of further instrumental reasoning for state-justified intentions and object-justified intentions can be taken for granted – no further argument is necessary. Nor does this reasoning internally permit an irrationality pill to be taken: in aiming at an action the agent will be disposed to do what needs to be done for that end.

The difference between these two solutions reflects a difference between two lines of thought that have been deployed elsewhere in this thesis: firstly, that an agent who aims at an action cannot have a coherent self-conception if they also have certain other intentions or beliefs; and secondly, that to intend is to exercise one's powers of agency in a certain direction in the way characteristic of action. The former, applied to instrumental reasoning, yields the view that unless an agent acquires further instrumental intentions, they will suffer an incoherent self-conception, and this is the origin of the power of state-justified intentions to create derivative instrumental intentions. The second line of thought yields the view outlined here: that what is justified is a stance of the agent towards an action they aim at, and this stance itself includes the usual range of instrumental capacities. Though both underlying lines of thought have a place, as I have argued, in an adequate theory of intention, it is the latter that is necessary to solve this problem.

Bibliography

Alonso, Facundo M. (2017). "Intending, Settling and Relying", in Shoemaker, David (ed.) *Oxford Studies in Agency and Responsibility Volume 4*, Oxford: Oxford University Press.

Archer, Sophie (2013). "Nondoxasticism About Self-Deception", *dialectica*, Vol.67 No.3: 265-282. https://doi.org/10.1111/1746-8361.12030

Aristotle (2002). *Nicomachean Ethics* (eds. Broadie, S. and Rowe, C, in *Nicomachean Ethics: Introduction, Translation and Commentary*). Oxford: Oxford University Press

Alvarez, Maria (2010). *Kinds of Reasons: Essays in the Philosophy of Action*. Oxford: Oxford University Press.

Andreou, Chrisoula (2008). "The Newxin Puzzle", *Philosophical Studies*, Vol. 139, No. 3: 415-422. http://www.jstor.org/stable/2773421

Anscombe, Elizabeth (1967). Who is Wronged? *The Oxford Review*, No. 5: Trinity

——. (1995). "Practical Inference", in Hursthouse R., Lawrence G. and Quinn W. (eds.) *Virtues and Reasons: Phillipa Foot and Moral Theory*, pp.1-34, Oxford: Clarendon Press.

——. (2001). *Intention* (2nd edition). Cambridge, MA: Harvard University Press.

Ashford, Elizabeth (2000). "Utilitarianism, Integrity and Partiality", *The Journal of Philosophy*, Vol. 97 No. 8: 421-439. https://www.jstor.org/stable/2678423.

Audi, Robert (2004). *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton, NJ: Princeton University Press.

Baier, Annette (1970). "Act and Intent", *The Journal of Philosophy*, Vo.67, No.19: 648-658

——. (1972). “Ways and Means”, *Canadian Journal of Philosophy*, Vol.1 No.3

——. (1985). *Postures of Mind: Essays on Mind and Morals*. Minnesota: University of Minnesota Press.

Bales, Adam (2020). “Intentions and Instability: a defence of causal decision theory”, *Philosophical Studies*, 177: 793-804, https://doi.org/10.1007/s11098-018-1206-4

Bratman, Michael (1984). “Two Faces of Intention”, *The Philosophical Review*, Vol. 93, No. 3: 375-405. https://www.jstor.org/stable/2184542.

——. (1987). *Intention, Plans and Practical Reason*. Center for the Study of Language and Information.

——. (1991). “Cognitivism About Practical Reason”, *Ethics*, Vol.102 No.1: 117-128. http://www.jstor.org/stable/2381726

——. (1998). “Toxin, Temptation, and the Stability of Intention”, in Coleman, Jules L., Morris, Christopher W. and Kavka, Gregory S. (eds.), *Rational Commitment and Social Justice: Essays for Gregory Kavka*, pp.59-83. Cambridge: Cambridge University Press.

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

——. (2009a). “Intention Rationality”, *Philosophical Explorations*, 12:3: 227-241. https://doi.org/10.1080/13869790903067717.

——. (2009b). “Intention, Practical Rationality and Self-Governance”, *Ethics*, Vol.119, No. 3:411-443. http://www.jstor.org/stable/10.1086/599985

——. (2014). “Temptation and the Agent’s Standpoint”, *Inquiry*, Vol 57 No.3. https://doi.org/10.1080/0020174X.2014.894271

Broome, John (2013). *Rationality Through Reasoning*. Chichester: Wiley-Blackwell.

——. (2015). “Responses to Setiya, Hussain and Horty”, *Philosophy and Phenomenological Research*, Vol. XCI No. 1. https://doi.org/10.1111/phpr.12205

Brunero, John (2005). “Two Approaches to Instrumental Rationality and Belief Consistency”, *Journal of Ethics and Social Philosophy*, Vol.1 No.1

Chan, David (1995). “Non-Intentional Actions”, *American Philosophical Quarterly*, Vol. 32, No. 2: 139-151. https://www.jstor.org/stable/20009814.

Chisholm, Robert (1970). “The Structure of Intention”, *The Journal of Philosophy*, Vol.67 No.19: 633-647. https://www.jstor.org/stable/2024584

Clark, Phillip (2001). “The Action as Conclusion”, *Canadian Journal of Philosophy*, Vol. 31 No.4: 481-505. https://www.jstor.org/stable/40232130

Cleveland, Timothy (1992). “Trying Without Willing”, *Australasian Journal of Philosophy*, Vol.70 No.3. https://doi.org/10.1080/00048409212345211

Cullity, Garrett and Gaut, Berys (eds.) (1997). *Ethics and Practical Reason*. Oxford: Oxford University Press.

Dancy, Jonathan (1993). *Moral Reasons*. Oxford: Blackwell.

——. (2018). *Practical Shape: A Theory of Practical Reasoning*. Oxford: Oxford University Press.

Danielson, Peter (ed.) (1998). *Modelling Rationality, Morality and Evolution*. Oxford: Oxford University Press.

Davidson, Donald (2001). *Essays on Actions and Events*, Oxford: Oxford University Press.

Dill, Brendan and Holton, Richard (2014). “The addict in us all”, *Frontiers in Psychiatry*, Vol. 5. https://doi.org/10.3389/fpsyt.2014.00139

Deigh, John (1996). *The Sources of Moral Agency: Essays in Moral Psychology and Freudian Theory*. Cambridge: Cambridge University Press.

Engstrom, Stephen (2009). *The Form of Practical Knowledge: A Study of the Categorical Imperative*. Cambridge, MA: Harvard University Press.

Fernandez, Patricio (2016). "Practical Reasoning: Where the Action Is", *Ethics*, Vol.126: 869-900.

Finlay, Stephen (2014). *A Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

Fix, Jeremy (2018). "Intellectual Isolation", *Mind*, Vol.127. https://doi.org/10.1093/mind/fzx046

Foot, Phillipa (1972). "Morality as a System of Hypothetical Imperatives", *The Philosophical Review*, Vol.81 No.3: 305-316. http://www.jstor.org/stable/2184328

——. (2002). *Virtues and Vices and Other Essays in Moral Philosophy*. Oxford: Clarendon Press.

Ford, Anton, Hornsby, Jennifer and Stoutland, Frederick (eds.) (2011). *Essays on Anscombe's Intention*. Cambridge, MA: Harvard University Press.

Ford, Anton (2016). "On What is in Front of Your Nose", *Philosophical Topics*, Vol.44 No.1:141-161

——. (2018). "The Province of Human Agency", *Noûs,* Vol. 52 No.3: 697-720. https://doi.org/10.1111/nous.12178

Frankfurt, Harry G. (1988). *The Importance of What we Care About*. Cambridge: Cambridge University Press.

——. (1999). *Necessity, Volition and Love*. Cambridge: Cambridge University Press.

Gallie, W.B. (1955-1956). “Essentially Contested Concepts”, *Proceedings of the Aristotelian Society*, Vol. 56: 167-198. http://www.jstor.org/stable/4544562

Gauthier, David (1963). *Practical Reasoning: The Structure and Foundations of Prudential and Moral Arguments and their Exemplification in Discourse*. Oxford: Oxford University Press.

——. (1984). “Deterrence, Maximization and Rationality”, *Ethics*, Vol. 94 No.3: 474-495. http://www.jstor.org/stable/2380819

——. (1986). *Morals by Agreement*. Oxford: Oxford University Press

Gibbard, Alan (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——. (2005). “Truth and Correct Belief”, *Philosophical Issues*, Vol. 15 No. 1: 338-350.

——. (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Harcourt, Edward (ed.) (2000). *Morality, Reflection and Ideology*. Oxford: Oxford University Press.

Hare, R.M. (1971). *Practical Inferences*. London: Macmillan

——. (1981). *Moral Thinking: Its Levels, Method and Point*. Oxford: Oxford University Press.

Harman, Gilbert (1976). “Practical Reasoning”, *The Review of Metaphysics*, Vol. 29 No.3: 431-463. http://www.jstor.org/stable/20126812

——. (1986). *A Change in View*. Boston, MA: MIT Press.

——. (2006). “Intending, Intention, Intent, Intentional Action, and Acting Intentionally: Comments on Knobe and Burra”, *Journal of Cognition and Culture* 6 (1-2): 269-276.

Hedden, Brian (2012). “Options and the subjective ought”, Philosophical Studies, Vol. 158: 343-360. https://doi.org/10.1007/s11098-012-9880-0

Herdova, Marcela (2018). “Trigger Warning: no proximal intentions required for intentional action”, *Philosophical Explorations*, 21:3: 364-383. https://doi.org/10.1080/13869795.2018.1435822.

Heuer, Ulrike (2010). “Reasons and Impossibility”, *Philosophical Studies*, Vol. 147: 235-246. https://doi.org/10.1007/s11098-008-9285-2

——. (2014). “Intentions and the Reasons for which we Act”, *Proceedings of the Aristotelian Society*, Vol. CXIV Part 3. https://doi.org/10.1111/j.1467-9264.2014.00374.x

——. (2018). “Reasons to Intend”, in Star, Daniel (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.

Hieronymi, Pamela (2006). “Controlling Attitudes”, *Pacific Philosophical Quarterly*, Vol. 87: 45-74.

——. (2009). “The Will as Reason”, *Philosophical Perspectives*, Vol.23.

——. (2011). “Reasons for Action”, *Proceedings of the Aristotelian Society*, Vol. 111: 407-427. http://www.jstor.org/stable/41331558

——. (2013). “The Use of Reasons in Thought (and the Use of Earmarks in Arguments”, *Ethics*, Vol. 124 No.1: 114-127. http://www.jstor.org/stable/10.1086/671402

Hillel-Ruben, David (2016). “A Conditional Theory of Trying.” *Philosophical Studies*, 173: 271–287. https://doi.org/10.1007/s11098-015-0490-5.

Hobbes, Thomas (1994). *Leviathan*. Indianapolis: Hackett Publishing Company

Holton, Richard (1999). “Intention and Weakness of Will”, *The Journal of Philosophy*, Vol. 96 No.5: 241-262. https://www.jstor.org/stable/2564667

——. (2004). “Rational Resolve”, *The Philosophical Review*, Vol.113 No.4: 507-535. https://www.jstor.org/stable/4148000

——. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

Hollis, Martin and Sugden, Robert (1993). "Rationality in Action", *Mind*, Vol. 102 No. 405: 1-35. https://www.jstor.org/stable/2254170

Hornsby, Jennifer (1980). *Actions*. London: Routledge and Kegan Paul.

——. (1995). "Reasons For Trying." *Journal of Philosophical Research*, XX: 525-538.

——. (2010) "Trying to Act", in Sandis, Constantine and O'Connor, Timothy. *A Companion to the Philosophy of Action*. Chichester: Wiley-Blackwell.

——. (2012). "Actions and Activity", *Philosophical Issues*, Vol. 22.

——. (2016). "Intending, Knowing How, Infinitives", *Canadian Journal of Philosophy*, Vol. 46 No.1: 1-17. https://doi.org/10.1080/00455091.2015.1132544

Horty, John F (2012). *Reasons as Defaults*. Oxford: Oxford University Press.

——. (2015). "Requirements, Oughts, Intentions", *Philosophy and Phenomenological Research*, Vol. XCI No. 1. https://doi.org/10.1111/phpr.12204

Hu, Jiajun (2017). "Rethinking the Videogame Case: Trying and Intending", *Philosophical Explorations*, Vol. 20 No. 3: 338-351. https://doi.org/10.1080/13869795.2017.1352017

Hubbs, Graham (2013). "How Reasons Bear on Intentions", *Ethics*, Vol. 124 No. 1: 84-100. https://www.jstor.org/stable/10.1086/671388

Hume, David (2009). *A Treatise of Human Nature* (eds. Norton and Norton). Oxford: Oxford University Press

Hurley, S.L. (1989). *Natural Reasons: Personality and Polity*. Oxford: Oxford University Press.

Hyman, John (2015). *Action, Knowledge and the Will*. Oxford: Oxford University Press.

Hursthouse, Rosalind (1991). “Arational Action”, *The Journal of Philosophy*, Vol. 88, No. 2: 57-68. https://www.jstor.org/stable/2026906

Jackson, Elizabeth (2018). “Belief and Credence: why the attitude-type matters”, *Philosophical Studies*, Vol.176:2477-2496. https://doi.org/10.1007/s11098-018-1136-1

Jackson, Frank and Pargetter, Robert (1986). “Oughts, Options and Actualism”, *The Philosophical Review*, Vol. 96 No. 2: 233-255. https://www.jstor.org/stable/2185591

Kavka, Gregory S (1978). “Some Paradoxes of Deterrence”, *The Journal of Philosophy*, Vol. 75 No. 6: 285-302. https://www.jstor.org/stable/2025707

——. (1983). “The Toxin Puzzle”, *Analysis*, Vol. 43 No. 1: 33-36. https://www.jstor.org/stable/3327802

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

Kiesewetter, Benjamin (2015). “Instrumental Normativity: In Defence of the Transmission Principle”, *Ethics*, Vol. 125: 921-946

——. (2017). *The Normativity of Rationality*. Oxford: Oxford University Press.

Knobe, Joshua (2003). “Intentional action and side effects in ordinary language”, *Analysis* 63:3:190-194.

Kolnai, Aurel (1961-1962). “Deliberation is of Ends”, *Proceedings of the Aristotelian Society*, Vol. 62: 195-218. https://www.jstor.org/stable/4544663

Kolodny, Niko (2005). “Why be Rational?”, Mind, Vol. 114 No. 455: 509-563. https://www.jstor.org/stable/3489006

——. (2010). “Ifs and Oughts”, *The Journal of Philosophy*, Vol. 107 No. 3: 115-143. https://www.jstor.org/stable/25700490

——. (2018). “Instrumental Reasons”, in Star, Daniel (ed.) *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press.

Körner, Stephen (ed.) (1974). Practical Reason: Papers and Discussions. Oxford: Blackwell.

Korsgaard, Christine (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

——. (2008). “Acting for a Reason”, in *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford University Press.

——. (2009). *Self-Constitution: Agency, Identity and Integrity*. Oxford: Oxford University Press.

——. (2019). “Constitutivism and the Virtues”, *Philosophical Explorations*, Vol. 22 No. 2: 98-116. https://doi.org/10.1080/13869795.2019.1599050

Lavin, Douglas (2011). “Problems of Intellectualism: Raz on Reason and its Objects”, *Jurisprudence*, Vol. 2 No. 2: 367-378 https://doi.org/10.5235/204033211798716817

——. (2013). “Must There be Basic Action?” *Noûs*, Vol. 47 No. 2: 273-301.

——. “Action as a Form of Temporal Unity: on Anscombe’s *Intention*”, *Canadian Journal of Philosophy*, Vol. 45: 609-629. https://doi.org/10.1080/00455091.2015.1125589

——. (2017). “Forms of Rational Agency”, *Royal Institute of Philosophy Supplement, Vol.* 80:171-193.

Lewis, David (1981). “Why ain’cha rich?”, *Noûs*, Vol. 15: 377-380.

Levy, Ken (2009). “On the Rationalist Solution to Gregory Kavka’s Toxin Puzzle”, *Pacific Philosophical Quarterly*, Vol. 90: 267-289.

Lord, Errol (2018). *The Importance of Being Rational*. Oxford: Oxford University Press.

Lovibond, Sabina and Williams, S.G. (eds.) (2000). *Identity, Truth and Value: essays for David Wiggins*. Oxford: Blackwell.

Marušić, Berislav and Schwenkler, John (2018). "Intending is Believing: A Defence of Strong Cognitivism", *Analytic Philosophy*, Vol. 59 No. 3: 309-340.

McDermott, Michael (2008). "Are Plans Necessary?", *Philosophical Studies*, Vol. 138 No. 2: 225-232. https://www.jstor.org/stable/40208870

McDowell, John (2010). "What is the Content of an Intention in Action?", *Ratio*, Vol. XXIII

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

——. (2004). "Anscombe on 'Practical Knowledge'", *Royal Institute of Philosophy Supplement*, Vol. 55: 43-68.

Mylonaki, Evgenia (2017). "Action as the Conclusion of Practical Reasoning; The Critique of a Rödlian Account", *European Journal of Philosophy*, Vol. 26 No. 1: 30-45. https://doi.org /10.1111/ejop.12175

Nagel, Thomas (1970). *The Possibility of Altruisim*. Oxford: Clarendon Press.

——. (1986). *The View From Nowhere*, Oxford: Oxford University Press.

Nguyen, C. Thi (2020). *Games: Agency as Art*. New York: Oxford University Press.

O'Brien, Lucy and Soteriou, Matthew (eds.) (2009). *Mental Actions*. Oxford: Oxford University Press.

O'Shaughnessy, Brian (1973). "Trying (as the Mental 'Pineal Gland')." *The Journal of Philosophy*, Vol.70 No.13, On Trying and Intending, 365-386. https://www.jstor.org/stable/2024676

——. (1980a). *The Will: A Dual Aspect Theory* (vol. 1). Cambridge: Cambridge University Press.

——. (1980b). *The Will: A Dual Aspect Theory* (vol. 2). Cambridge: Cambridge University Press.

Paul, Sarah K. (2013). "The Conclusion of Practical Reasoning: the Shadow between Idea and Act", *Canadian Journal of Philosophy*, Vol. 43 No. 3: 287-302. https://doi.org/10.1080/00455091.2013.847345

——. (2014). "Diachronic Incontinence is a Problem in Moral Philosophy", *Inquiry*, Vol. 57 No. 3: 337-355. https://doi.org/10.1080/0020174X.2014.894273

Parfit, Derek (2001). "Rationality and Reasons" in Egonsson, Jonas Josefsson, Björn Petersson and Toni Rønnow-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values*, pp. 19-39, Alderhot: Ashgate.

Pears, David (1998). *Motivated Irrationality*. Oxford: Clarendon Press.

Pink, Thomas (1996). *The Psychology of Freedom*. Cambridge: Cambridge University Press.

——. (2016). *Self-Determination: The Ethics of Action, Volume 1*. Oxford: Oxford University Press.

Price, A.W. (2004). "On the So-Called Logic of Practical Inference", *Royal Institute of Philosophy Supplement* Vol. 54:119-140.

——. (2008). *Contextuality in Practical Reason*. Oxford: Clarendon Press.

Quinn, Warren (1980). "Putting Rationality in its Place", in Frey and Morris (eds.) (1993), *Value, Welfare and Morality*, pp.26-50. Cambridge: Cambridge University Press.

Raz, Joseph (1999). *Engaging Reason: on the Theory of Value and Action*. Oxford: Oxford University Press.

——. (2011). *From Normativity to Responsibility*, Oxford: Oxford University Press.

Rinard, Susanna (2018). "Believing for Practical Reasons", *Noûs*. https://doi.org/10.1111/nous.12253

Rorty, Amélie (ed.) (1980). *Essays on Aristotle's Ethics*. Berkeley, CA: University of California Press.

Ross. W.D. (1930). *The Right and the Good*. Oxford: Clarendon Press

Russell, Devlin (2017). "Intention as Action under Development: why Intention is not a Mental State", *Canadian Journal of Philosophy*. https://doi.org/10.1080/00455091.2017.1414524

Ryle, Gilbert (1963). *The Concept of Mind*, Harmondsworth: Penguin Books.

Scanlon, Thomas (1998). *What we Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

——. (2014). *Being Realistic About Reasons*. Oxford: Oxford University Press.

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

——. (2012). "The Ubiquity of State-Given Reasons", *Ethics*, Vol. 122 No. 3: 457-488. https://www.jstor.org/stable/10.1086/664753

——. (2013). "State-Given Reasons: Prevalent, if not Ubiquitous", *Ethics*, Vol. 124 No. 1: 128-140. https://www.jstor.org/stable/10.1086/671385

Schroeder, Severin (2001). "The Concept of Trying", *Philosophical Investigations*, 24:3: 213-227.

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

Shah, Nishi (2003). “How Truth Governs Belief”, *The Philosophical Review*, Vol. 112 No. 4.

——. (2008). “How Action Governs Intention”, *Philosophers’ Imprint*, Vol.8, No.5: 1-19. www.philosophersimprint.org/008005.

Shah, Nishi and Silverstein, Matthew (2013). “Reasoning in Stages”, *Ethics*, Vol. 124 No. 1: 101-113. https://www.jstor.org/stable/10.1086/671387

Sidgwick, Henry (1962). *The Methods of Ethics (7th edition)*. London: Macmillan

Sinhababu, Neil (2013). “The Desire-Belief Account of Intention Explains Everything”, *Noûs*, Vol. 47 No. 4: 680-696. https://doi.org/10.1111/j.1468-0068.2012.00864.x

Skorupski, John (2010). *The Domain of Reasons*. Oxford: Oxford University Press

Snedegar, Justin (2017). “Time-Slice Rationality and Filling in Plans”, *Analysis Reviews*, Vol. 77 No. 3: 595-607. https://doi.org /10.1093/analys/anx070

Sorensen, Roy (1988). *Blindspots*. Oxford: Clarendon Press.

Steward, Helen (1997). *The Ontology of Mind: Events, Processes and States*. Oxford: Clarendon Press.

——. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Stout, Rowland (1996). *Things that Happen Because They Should: A Teleological Approach to Action*. Oxford: Oxford University Press.

——. (2019). “Practical Reasoning and Practical Knowledge”, *Canadian Journal of Philosophy*, Vol. 49 No. 4: 564-579. https://doi.org/10.1080/00455091.2018.1463839

Tenenbaum, Sergio (2007a). “The Conclusion of Practical Reason”, in Tenenbaum, Sergio (ed.) *New Trends in Philosophy: Moral Psychology,* pp. 323-343. Rodopi.

——. (2007b). *Appearances of the Good: An Essay on the Nature of Practical Reason*. Cambridge: Cambridge University Press.

——. (ed.) (2010). *Desire, Practical Reason and the Good*. Oxford: Oxford University Press.

——. (2014) “Minimalism About Intention: A Modest Defence”, *Inquiry*, Vol. 57 No. 3: 384-411. http://dx.doi.org/10.1080/0020174X.2014.894275

——. (2018). “Reconsidering Intentions”, *Noûs*, Vol. 52 No. 2: 443-472. http://doi.org/10.1111/nous.12160

——. (2020). “The Guise of the Good”, in Chang, Ruth and Sylvan, Kurt (eds.) *The Routledge Handbook of Practical Reason*, London: Routledge.

Thompson, Michael (2008). *Life and Action: Elementary Structures of Practical Thought*. Cambridge, MA: Harvard University Press.

Velleman, David (1985). “Practical Reflection”, *The Philosophical Review*, Vol. 94, No. 1: 33-61. http://www.jstor.org/stable/2184714.

——. (1991). Review of *Intentions, Plans and Practical Reason* by Michael E. Bratman, *The Philosophical Review*, Vol. 100 No. 2: 277-284. https://www.jstor.org/stable/2185304

Verbeek, Bruno (2014). “On the Normativity of Intentions”, *Topoi*, Vol. 33: 87-101. http://doi.org/10.1007/s11245-013-9221-8

Wallace, R. Jay (2001). “Normativity, Commitment and Instrumental Reason”, *Philosophers’ Imprint*, Vol. 1 No. 3. http://www.philosophersimprint.org/001003

——. (2006). *Normativity and the Will: Selected Papers on Moral Psychology and Practical Reason*. Oxford: Clarendon Press.

Wedgwood, Ralph (2017). *The Value of Rationality*. Oxford: Oxford University Press.

Wiggins, David (1987). *Needs, Values, Truth: Essays in the Philosophy of Value*. Oxford: Blackwell.

——. (2006). *Ethics: Twelve Lectures on the Philosophy of Morality*. London: Penguin.

Williams, Bernard Arthur Owen (1973). *Problems of the Self*. Cambridge: Cambridge University Press

——. (1981). *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press.

——. (1993). "Moral Incapacity", *Proceedings of the Aristotelian Society*, Vol. 93: 59-70. https://www.jstor.org/stable/4545165

——. (1995). *Making Sense of Humanity and other philosophical papers, 1982-1993*. Cambridge: Cambridge University Press.

——. (2008). *Shame and Necessity*. Berkeley, CA: University of California Press.

Winch, Peter (1972). *Ethics and Action*. London: Routledge and Kegan Paul.