



Action, Reason and Will

Pascal Hussein Ally

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Word Count (excluding bibliography): 71, 420

The University of Sheffield
Faculty of Arts and Humanities
School of History, Philosophy and Digital Humanities

March 2025

Table of Contents

<i>Acknowledgements</i>	<i>ii</i>
<i>Abstract</i>	<i>iv</i>
Introduction	1
Chapter One: Understanding Human Agency: Action and Agent	4
Chapter Two: On Human Agency as Planning Agency: Three Problems for Bratman's Planning Theory of Intention and Action	32
Chapter Three: Agency as a Way of Knowing? Velleman on Action, Intention and Self- Knowledge	58
Chapter Four: Agency Within Morality? Korsgaard on the Practical Standpoint, Free Will and the Moral Law	86
Chapter Five: Action, Reason and Will: Consciousness, Judgement, Commitment	112
Chapter Six: Action, Reason and Will: Intention Sublates Judgement and Desire	135
Conclusion	160
Bibliography	162

Acknowledgements

The pursuit of this PhD and the completion of this dissertation would not have been possible without the various forms of support and encouragement that I have been fortunate to receive from different groups and individuals. I would like to register my gratitude to them here.

I would like to thank first the Faculty of Arts and Humanities at the University of Sheffield for awarding me a full doctoral scholarship. Without this generous financial support I would not have taken up my PhD admission offer and come to Sheffield.

Next, my profound gratitude and appreciation go to my doctoral supervisors, James Lenman and Max Khan Hayward, for guiding me with patience, solicitude and understanding through the development and completion of this thesis. The supervision meetings and other discussions that I had with them were enjoyable, illuminating, and helped keep me on track. I am grateful, in particular, to Jimmy, my primary supervisor, who met with me once (and sometimes twice) every month to discuss my work and who read and commented on the different drafts of each chapter. Jimmy was incredibly accommodating and generous with his time and his academic wisdom during the entire course of my PhD.

I am also deeply thankful to the School of History, Philosophy and Digital Humanities for its exceptionally supportive and intellectually stimulating environment. I received a tremendous amount of support to navigate welfare challenges and administrative demands during my doctoral journey. For this I thank especially Ed Matthews, who was the PGR and Research Support Officer for Philosophy during much of my PhD. Ed was a kind and knowledgeable angel in human form every time I needed help with administrative procedures. I also benefited from the engaging discussions I had with, and the helpful remarks I received from, audiences at the Department Graduate Seminar, where I presented parts of this thesis over the four years of my PhD. The Seminar was such a congenial and exciting research-sharing environment, and our discussions typically continued over drinks at the University Arms. For such excellent philosophical exchanges I thank particularly George Surtees, Alvaro Rodriguez-Gonzalez and Charlie Potter.

Further, I would like to thank Carien Smith, Rod Howlett (who offered to proofread the entire thesis), Aishwarya Suresh, Srijan Kumar Jha, Isuru Nanayakkara, Kajanan Selvaranjan and Émile Motanda. Their friendship, generosity and support made the whole doctoral journey smoother.

Last but not least, very special thanks are due to the team of nurses at the Headache Clinic in the Neurology Department of the Royal Hallamshire Hospital in Sheffield: with an unwavering devotion and kindness they battled the recalcitrant monster in my head to enable me to pursue this PhD to its completion.

Abstract

This thesis investigates the constitution of human action. What we do depends on what we believe and what we want. Or our actions depend on our reason and our will. Understanding human action thus requires grasping the nature of human reason and will and the relationship between them. In this thesis, I investigate the constitution of human action both by examining some contemporary accounts of it and by developing an alternative account. After defending a concept of human action as intentional action by a self-governing agent (Chapter One), I discuss three different accounts of human agency with which I share a dissatisfaction with the desire-belief model. Chapter Two explores Bratman's planning theory, which extends the desire-belief model with the addition of intentions construed as elements of coordinating plans and as rational constraints on practical reasoning. I raise some problems for the conception of intention in terms of plans and the function of coordination. Chapter Three scrutinises Velleman's account, which adds to the desire-belief model the conception of agency as a special way of knowing and the notion of intention as a cognitive attitude that embodies self-knowledge. I argue, *inter alia*, that this intellectualist account of agency seems inconsistent with genuine self-determination. Chapter Four appraises Korsgaard's account, on which we constitute ourselves into agents by choosing our actions in accord with the moral law. I show, *inter alia*, the impossibility of a logical derivation of the moral law from the premise of deliberative freedom. Finally, Chapters Five and Six develop an account of the constitution of human action, on which when we act intentionally, we act under the guise of the good, which we apprehend first through our emotions or desires and then through practical reason, and which we volitionally commit ourselves to bring about.

Introduction

What we do depends on what we believe and what we want; or our action depends on our reason and our will. This is a truism or a near truism. But it is not trivial. Making sense of it is one of the long-standing problems of philosophy. It requires grasping the nature of reason and the will, the relation between reason and the will in the constitution of our action, and even the nature of the human being. But each of these issues is chronically perplexing and contentious. Making sense of how our action depends on our reason and our will is what we might call *the problem of the constitution of human action*.

This thesis investigates the problem of the constitution of human action. My investigation of this problem involves both an examination of contemporary treatments of it and a development of my own novel account. This investigation is presented across six chapters.

Chapter One presents and develops a concept of human action that informs my engagement with the problem of the constitution of human action in the thesis. This is the concept of human action as intentional action by a self-governing agent. This chapter concludes with the observation that the standard (or desire-belief) theory of action does not capture human agency. According to this theory, when we act intentionally, we perform the action in the belief that it is a means to the fulfilment of a desire we happen to have. And the desire-belief pair rationalises the said action. I argue that the agent that this theory portrays is not recognisably human. From here, I proceed to examine (in Chapters Two through Four) three different accounts of the constitution of human action developed against the backdrop of the inadequacy of the desire-belief theory of action.

Chapter Two examines Bratman's account of human agency as planning agency. Bratman has made the phenomena of plans and planning the cornerstone of his account of intention, intentional action and practical rationality. In his influential *Intention, Plans, and Practical Reason* (1987), he argues that the standard (or desire-belief) theory of intentional action needs a substantial extension (in both its descriptive and normative aspects) to make room for the sort of agents that we are: planning agents, or agents who need plans for intra-personal and interpersonal coordination. The extension of the descriptive aspect consists in recognising

intention as a distinctive mental state alongside desire and belief, and that of the normative aspect in allowing that intention has a direct relevance to the rationality of actions. In this chapter, I examine only Bratman's account of intention as an addition to the psychological or descriptive aspect of the desire-belief model, or of what sort of mental state it is. And I argue that Bratman's planning theory of intention and action faces three problems, which I discuss in the chapter.

Chapter Three discusses Velleman's account of human agency as 'a special way of knowing'. According to this account, in order to understand our agency we need to understand our self-knowledge. I dub this view *epistemicism about human agency*. In this chapter, I examine Velleman's epistemicism about agency through what we may take as its three constitutive theses. First, practical knowledge is 'a very different knowledge'. Call this thesis *exceptionalism about practical knowledge*. Second, practical knowledge is embodied in the agent's intention, which is a kind of belief or cognitive attitude. Call this view *cognitive intentionalism about practical knowledge*. And, third, acting autonomously is acting in accordance with one's representation or cognitive awareness of one's (strongest) motives. I argue that both exceptionalism and cognitive intentionalism are implausible, and that epistemicism about agency does not secure any genuine self-determination for human agents.

Chapter Four critically engages with Korsgaard's Kantian account of human agency as self-constitution. Korsgaard argues that, as rational creatures, we constitute ourselves into agents by choosing our actions in accordance with the categorical imperative or the moral law. Korsgaard claims that the moral law as the principle of choice for rational agents follows from the fact that freedom is a fundamental feature of the practical standpoint. Her argument for this position can be presented as follows: Freedom (from natural sensible inclinations) is a fundamental feature of the standpoint of practical reason and choice - that is, we must choose under the idea of freedom when we deliberate about what to do. If freedom (from natural sensible inclinations) is a fundamental feature of the standpoint of practical reason and choice - that is, if we must choose under the idea of freedom when we deliberate about what to do -, then the categorical imperative is the fundamental principle of choice for a free will. If the categorical imperative is the fundamental principle of choice for a free will, then a free choice is a moral choice. In this

chapter, I show the impossibility of a logical derivation of the moral law from the premise of deliberative freedom.

Finally, in Chapters 5 and 6, I attempt to provide and develop a novel account of the constitution of human action by tackling the issues of the nature of the human being and of the nature and relation of reason and will. In Chapter 5, I develop accounts of human consciousness, practical reason, and the will with a view to explaining the constitution of human action. I first argue that human consciousness is distinctively structured by an enquiry-guided pattern of four levels of conscious intentionality. The basic idea here is that our consciousness is structured in such a way that we are characteristically inquisitive about, and hence active with respect to, our perceptual and mental experiences. And this internal structure is critical for distinctively human agency because it makes possible the exercise of practical reason and the will. I then develop an account of practical reason as our capacity for evaluative thinking about what to do. I argue that practical reasoning concludes in judgements of value about what to do, and that these judgements are essentially constituted by feelings of value, factual judgements and imaginative judgements. Next, I sketch an account of the will as a distinctive capacity of the mind, and argue that the commitment involved in a state of intention is an existential commitment. And Chapter 6 develops an account of the relation between reason and will in the constitution of human action by arguing that this relation is most plausibly construed as one of sublation (i.e. dependence and transcendence).

Chapter One: Understanding Human Agency: Action and Agent

Philosophers have intended the concept of the will to distinguish action from mere bodily movement. The locus of responsibility then because the agent's identification with his actions through acts of will (Price, 2010, p. 29).

How do we, humans, make things happen? Suppose that your arm goes up during an auction at Auction House London. What makes it the case that you raised it (say, to enter a bid), as opposed to its being caused to go up by, say, a muscular spasm or a stimulation of your motor cortex with a micro-electrode, and so to its being something happening to you? To put it in an equivalent way, when your arm thus goes up, what makes it the case that you are an *agent* of this bodily movement as opposed to being a mere *patient* (Smith, 2014, p. 80)? Or what makes it the case that it is *you* who have initiated this event, rather than some sub-personal mechanism or some psychological subsystem operating independently inside you?

The concern of this chapter is not to address straight away these questions about the explanation of human action. That will be the business of the subsequent chapters. This chapter seeks instead to present and develop a concept of human action that will inform my engagement with these questions. It is the concept of human action as intentional action by a self-governing agent. Presenting and developing this concept involves defending the following constitutive claims: first, in line with these questions, I hold that a two-fold categorisation of human events - into actions and mere happenings - is adequate for the interests of action theory (Section 1). Second, I hold that understanding the difference between these events requires grasping first the nature of the agent. I further hold that the human agent is distinctively a self-governing agent and that a person's self-governing ability to do things is plausibly accounted for in agent-causal rather than event-causal terms (Section 2). And, third, in the question of what makes it the case that a human event is an action or a mere happening, I take it that the phenomenon of interest here is an intentional action. Further, I take it that an intentional action is an action done for reasons; that doing something intentionally involves intending to do it; and that there is no principled distinction between intentional action and self-governed action (Section 3).

Since these claims are important and controversial in the philosophy of action, I defend each of them in this chapter. And given the breadth of each of them, and the amount of space available in a chapter, my treatment of the claims will, sadly, not be detailed enough.

1

Let's begin with the categorisation of human events as the first step toward understanding the phenomenon of interest to us here. The questions asked at the opening of this chapter suggest that there are certain instances of our arm's going up that are things that we do and certain others that are things that happen to us. They contrast only two kinds of events in a human life: events that are actions and events that are mere happenings. This contrast is in the spirit of classical philosophy of action, which conceived of its central problem as that of explicating the difference between human events that are instances of being active and those that are instances of being passive. This problem is usually seen as being captured by Wittgenstein's famous question in *Philosophical Investigations*: 'What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?' (1958, §621).

In this section, I want to defend this classical twofold classification of human events against the charge by Frankfurt (1988) and Velleman (2000) that it is inadequate for the interests of the theory of action because it overlooks human events that are instances neither of actions nor of mere happenings. Examples of these events include idly and inattentively drumming one's fingers on the table, or idly and inattentively scratching one's head while engrossed in a book. For Frankfurt, events such as these are instances of being active or behaving mindlessly - like a spider 'when it moves its legs in making its way along the ground' - but 'without lapsing into mere passivity' (1988, p. 58). And Velleman characterises them as 'the things that one does [but] somehow fails to regulate in the [autonomous] manner' or as instances of 'a partial and imperfect exercise of the subject's capacity to make things happen (2000, p. 4). Velleman does not say what this autonomous regulation consists in or how the subject's capacity to make things happen is exercised fully and perfectly, but we might infer from his statement that this sort of human events are events that are deficient in something. Frankfurt hinted that these events are deficient in conscious attention. And Velleman describes them as lacking in self-regulation (or

control). For both Frankfurt and Velleman, these behaviours are mere activities, not actions and not mere happenings. So, for these philosophers, there are three kinds of human events.

The phenomena that Frankfurt and Velleman identify as members of the class of mere activities have a long history in philosophical discussions of human agency. One of the earliest mentions is perhaps by Aquinas. In the second part of his *Summa Theologiae*, he enquires into the status of the many things that we do without deliberation, or even without thinking - for example, when someone absent-mindedly rubs his chin. He calls these behaviours 'acts of man' and he contrasts them with 'human acts,' the latter being the only ones that are characteristically ours *qua* human beings or creatures endowed with reason and will. Human acts are those actions that lie under man's control or which issue from his free decision as an ability of reason and will. Acts of man, by contrast, are behaviours that 'do not come from reasoned deliberation, which sets the stage on which men act as men' (Ia2ae, q1, a1). This category includes the absent-minded rubbing of one's chin and the occurrences identified by Frankfurt and Velleman. So Aquinas would probably agree with such behaviours being called 'mere activities' and not actions, though for a rather different reason, namely, that they do not issue from episodes of practical reasoning. And we can see in Aquinas's contrast between 'acts of man' and 'human acts' a precursor of Frankfurt's and Velleman's distinction between 'mere activity' and 'action'.

The constitution of the category of mere activities (as behaviours essentially deficient in conscious attention and control) suggests what we may call *the problem of consciousness and control threshold for agency*. It is natural to think that consciousness and control are necessary for agency, or at least for human agency. Intuitively, an occurrence cannot be attributed to a person as his action if it escapes his consciousness and control. But, as consciousness (awareness, attention, etc.) and control come in degrees, it is not clear how much of them is required or to what extent they must be exercised by a person in order for him to count as acting. It is not clear, in other words, how and where to set the threshold for consciousness and control. Tentatively, we may say that the concept of consciousness and control threshold is not meaningfully applicable to mere happenings but it is to mere activities and to actions, with the main difference being, presumably, that mere activities occur at a sub-threshold level or do not reach the (yet undefined) fixed threshold for human agency. And so the introduction of the

intermediate category of mere behaviours (as sub-actions or infra-actions) forces us to confront the rather difficult question of how much conscious attention and control an action requires.

But to examine the validity of the category of mere activities, we do not need to settle this question. In my view, to examine it we only need to consider whether the so-called mere activities have anything (important) in common with a particular subset of highly valued actions - 'skilled performances'. A comparison of this sort will show that the distinction between mere activities and actions - at least as drawn by Frankfurt and Velleman but also as it can be derived from Aquinas - is not valid. It will show that their cases of mere activities are instances of actions.

My argument is that what is presented as the distinctive feature of mere activities, namely, deficiency in conscious attention, awareness and control, is also a significant property of a subset of behaviours - 'skilled performances' - whose status as full human actions is never in doubt for us. If this is the case, then the supposed distinctive quality of mere activities fails to set them apart from actions. In what follows, I first argue that the so-called mere activities are human actions because their supposed distinctive property is an essential property of skilled human actions (1.1). I then examine Velleman's treatment of the Freudian slip of the tongue as a paradigmatic case of merely motivated activity, and argue that slips of the tongue, like slips of the finger, are perhaps best understood as actions that are unintentional under some descriptions. I conclude this section with brief remarks on the possibility of borderline cases - such as sleepwalking behaviours - (1.3). The upshot of my argument in this section is that, admitting of some borderline cases, Wittgenstein's twofold typology of human events is sufficiently adequate for the purposes of action theory.

1.1. To begin with, consider skilled performance in commonplace activities such as driving a car, riding a bike, hammering nails and splitting firewood, and in elite activities such as playing tennis, playing violin and ballet dancing. Let's call the former kind 'commonplace skilled performance' and the latter 'elite skilled performance'¹. There is no doubt, I think, that driving a car, chopping wood, playing tennis and ballet-dancing are actions - and, in a sense, distinctively

¹ The phrases 'commonplace skilled performance' and 'elite skilled performance' are my analogues for Bergamin's (2017) phrases 'everyday expertise' and 'esoteric expertise'.

human actions. And, obviously, the same goes for skilled car-driving, skilled wood-chopping, skilled tennis-playing and skilled ballet-dancing. Now consider the so-called mere behaviours such as absently rubbing one's chin and drumming one's fingers on the table idly and inattentively. We are told that these behaviours fail to qualify as actions because they are deficient in conscious awareness and control. If so, actions - and particularly skilled actions of both sorts - must be markedly different in this regard. Interestingly, however, a property that is common, and even essential, to both kinds of skilled performances appears to be shared also by the so-called mere activities: deficiency in conscious attention and control.

Research on skill and automaticity reveals that 'as expertise in some domain increases, the need to focus conscious attention on certain segments of actions in that domain decreases' (Shepherd, 2015, pp. 328-9). These segments may have required conscious attention at one time during the process of learning or skill development. But once they have been over-learned and made automatic by dint of practice, conscious attention becomes superfluous for their performance. What may require conscious or supervisory attention in any given (skilled) action is likely to be only those aspects that are novel, difficult or especially significant (Baars, 1992, p. 7). So a skilled driver can change gears absent-mindedly or in a state of peripheral awareness of the movements of his feet or while fully absorbed in the music playing on the radio. There is no need for him to be consciously or cognitively attentive to his driving or the car gears. Because he is skilled, he can drive or shift gears automatically, with his mind on something else.

Indeed, an essential part of being skilled at activities of both (commonplace and elite) kinds is that we can and typically do perform them 'in the flow' with 'smooth coping,' as Dreyfus puts it - that is, without any need for explicit thought or conscious attention (Bergamin, 2017). The skilled driver may thus need to exercise conscious attention and control only at certain points or places on his journey, or the skilled classical violinist only when starting a technically difficult movement of a violin concerto. During the rest of their skilled performances, our driver and violinist can act with their minds out of what they are doing. That is, they can perform their actions automatically and without conscious control - much like Aquinas's individual absently rubbing his chin or Velleman's reader idly and inattentively scratching his head while engrossed in a book. And we would not say that in so behaving our skilled performers are not *acting* but have lapsed into *mere activity*.

It is easy to understand why the so-called mere activities deserve the status of human actions when we appreciate the fact that acting (intentionally) does not require as much conscious thinking and deliberation as philosophers of action generally assume. (Just consider the many things you do every day and how you go about doing them.) The reason for this seems to be that automatisms or habits are the building blocks of most - if not all - of our actions. And part of the reason for the centrality of habituation or automaticity is that our mind's processing resources are limited and get easily depleted. It is a fact, documented by psychologists, that our consciousness is very limited in its processing resources. Even simple activities like memorising someone's phone number, counting the pens and highlighters on your office table, making up your mind about whether to eat the cookie before you, or reading this sentence, consume a significant amount of your conscious processing resources (Hassin et. al., 2009, p. 549).

We can plausibly speculate that it is because of this handicap that our nature has made habituation central to our agency. Habituation makes it possible for us to shift from the consciously aware mode to the autopilot mode of goal-pursuit. As such, it is greatly beneficial to creatures with limited mental resources such as ourselves. William James (1890) identified and explained two main benefits of habituation. Firstly, it simplifies the movements required to achieve a given result, makes them more accurate and diminishes fatigue (pp. 117-8). Secondly, habituation diminishes the conscious attention with which our acts are performed. Once the performance of an action has become habitual, the occurrence of one movement calls up its appropriate successor movement and this one the next without the intervention of the conscious will and 'as if [the first movement] and the rest of the chain were fused into a continuous stream' (p. 119). James's view is that conscious attention plays a role early in the process and then becomes less necessary. And he saw part of the value of shifting from conscious to non-conscious acting in saving up limited mental energy (p. 126).

These remarks about habituation and mental energy saving are clearly relevant for our examination of those types of behaviours judged deficient in conscious attention and control, and hence non-actions, by Frankfurt and Velleman. These philosophers do not seem to have considered how limited our mental processing resources are or to have sufficiently appreciated the significance of habituation or automaticity for human agency. When we appreciate these

facts, it becomes easy to make sense of the so-called 'mere activities' and to vindicate them as human actions. In their defence, we can make these four final remarks. First, some of our actions require more conscious attention and effort than others. It does not follow from this fact that we are more 'active' in the former than in the latter. For example, reading a book obviously requires more conscious attention than scratching one's head, but this does not make the latter a lesser action or a non-action. It often costs people with certain neurological conditions a lot of mental and muscular energy to scratch their head when they need it. Second, because of our mental handicap, each one of our actions is bound to require less conscious attention and effort in the process of our becoming skilled at performing it. If such was not the case, many of our goals would go unpursued for lack or insufficiency of energy. Third, habituation and attention-switching are our natural systems for saving up our limited mental energy and thus enabling us to pursue our various goals at not too high a cost. Finally, certain actions are perhaps best, or ought to be, performed mindlessly: for example, chewing one's food or trying to fall asleep. You surely run the risk of failing to sleep when you consciously try to fall asleep.

1.2. At this point I would like to consider a particular type of behaviour that Velleman seems to treat as a paradigmatic case of merely motivated activity: the Freudian slip of the tongue. Think of President George H. W. Bush's famous slip of the tongue in his 1988 speech on television: 'We've had triumphs. Made some mistakes. We've had some sex... uh... setbacks.' Ascribing to Freud's interpretation of such slips, Velleman claims that, in one sense, they are produced by the subject; in another, they are produced despite the subject, by desires that he didn't intend to express (2000, p. 4). Now do Freudian slips support the case for a distinct category of human events? To find out, we may first want to know if there is anything in common between the case of Bush's slip of the tongue and the case of someone idly and inattentively scratching his head. The head-scratching individual seems to be 'regulating' his behaviour, which, by definition, does not happen in a slip of the tongue. As we have seen, Velleman characterises mere activities as 'partial and imperfect' exercises of the 'capacity to make things happen' (p. 4). While it is not obvious what is 'partial and imperfect' about the act of idly and inattentively scratching one's head while engrossed in a book, a slip of the tongue surely reveals some imperfection. But what kind of imperfection?

Presumably, slips reveal a mismatch between a reportable governing intention and its execution (Baars, 1992, p. 3). There is a mismatch between what President Bush intends to say and what he produces in trying to execute his intention. And this mismatch certainly results from a failure of voluntary control. But what explains or causes this failure? This, I think, is an empirical question for psychologists to investigate. It is a controversial issue among them whether verbal slips like President Bush's reveal, and are motivated by, unconscious desires or whether they are just trivial failures in the brain's workings. But whatever the truth of the matter, I want to suggest that slips of the tongue are best understood, at least from the agent's own perspective, as actions that are unintentional under certain descriptions. To see why, consider first the difference between a slip like President Bush's and its voluntary imitation (by the agent himself). The two seem to be quite identical behaviourally, but different only psychologically. One significant difference surely has to do with control: it fails in the case of the slip but succeeds in the case of its imitation.

Another, related difference pertains to intentionality. Perhaps this is best shown by examining an analogous case - that of a non-verbal slip: for instance, the slip of the finger. Consider this example by Zhu (2004, p. 189): suppose you are in front of an automatic hot beverage vendor intending to buy a cup of Colombian Decaf only to find, to your surprise, that you've actually pushed the button for Cappuccino. What's the difference between your pushing the Cappuccino button this way and President Bush's saying 'We've had some sex'? Isn't a slip of the finger the behavioural analogue of a slip of tongue? We want to say that your pushing the Cappuccino button was unintentional, that it was an unintentional action, and not that it is a mere activity motivated by your unconscious desire for Cappuccino. If we do not commonly suspect a desire operating anarchically behind a slip of the finger, why should we assume such a desire behind a slip of the tongue? One may object that it is not to folk psychology to adjudicate such cases, but to scientific psychology. Admittedly so. The more important point here, however, is that even if Freud's thesis that slips are creatures of unconscious desires was true, we have good reason not to classify behaviours as mere activities on the basis of the presence of unconscious motives in their causal history. Psychology has taught us that unconscious processes and mechanisms are so pervasive that few of our actions escape their influence.

1.3. To end this discussion, I would like to remark that my rejection of the category of mere behaviours should not be interpreted as a rejection of the possibility of borderline cases, cases that straddle the line between actions and mere happenings. I believe in the existence of such cases. Perhaps a particularly puzzling example of borderline cases is the behaviours exhibited by sleepwalkers. These behaviours can be simple or complex, innocuous or violent. Perhaps the most famous, or rather infamous, case of complex and violent sleepwalking behaviour is the case of (attempted) murder by a 23-year-old Canadian electronics worker named Kenneth Parks. In the early morning hours of May 24, 1987, Parks drove 14 miles to the house of his in-laws, struck his mother-in-law with a tire iron and stabbed her to death, and then choked his father-in-law almost to death. Shockingly, a year later, Park was cleared of the charges of murder and attempted murder on the ground that he was sleep-walking during the whole incident. And so he had not exercised any meaningful control over his behaviours.

Studies of the brain activity of sleepwalkers reveal that during an episode of sleepwalking the brain areas involved in motor control are highly active, whereas the brain regions responsible for higher cognitive functions (such as understanding, judgement, evaluation and planning) operate as if in deep sleep (Siclari et.al., 2012, pp. 40-41). If this is true, we may say that Parks physically controlled his behaviours (- a feature of action) but didn't mentally or intentionally control them (- a feature of a mere happening). In other words, he enjoyed executive control over his behaviours but lacked higher-order, reflective or autonomous control over them, and so his behaviours straddled the line between actions and mere happenings. This remark takes us straight into the terrain of the nature of the human agent in the following section.

2

In the preceding section, I argued for the adequacy of Wittgenstein's twofold classification of human events for the interests of the philosophy of action. This is the classification of human events, like an arm raising, into actions and mere happenings. And it raises the question of what account we can give of the difference between these two events - say, my arm going up as a result of a muscular contraction and my arm going up because I raised it. It is my contention that grasping and accounting for this difference requires understanding first the subject or agent. For one thing, events like actions or mere happenings exist only for creatures capable of agency, and

it is only such creatures that can experience an event as an action or as a happening. For another, agency, or the exercise of the capacity to act, is ‘a general phenomenon among living things of all kinds and quite generally across the natural world’ (Crittenden, 2012, p. 72). And so is, presumably, the contrast between actions and mere happenings. We may then say that ‘[w]hat any given agent [does or] is able to do is a function of what kind of being it is, or what powers it has’ (Crittenden, p. 72). Thus, to understand the behaviour of a certain agent, we need to know what sort of creature that agent is. To understand human action, which is the object of my concern in this thesis, we have to understand the ‘human agent’ or our ‘selves’.

That is why, in this section, I focus on trying to understand the human agent. Now what is the nature of the human agent? This is a complex or multifaceted question, one that requires far more space and analysis than I can afford here. But a common way to approach this question is to begin by stating that the human agent is - in a certain sense of self-governance - essentially and distinctively a self-governing agent. This is now a widely accepted view in the philosophy of action. A question that usually arises in this connection is whether a human person’s self-governing ability to do things is plausibly made sense of in event-causal or in agent-causal terms. That is, whether human agent causation is ontologically reducible to event causation.

In what follows, I try to illuminate the concept of human self-governance by drawing from certain aspects of our conscious experiences of motivation and acting, which I will call ‘gaps of activity’ or, more precisely, ‘gaps of rational and volitional activity’ (2.1). From this discussion of the phenomenology of our agency I move on to the ontology of agent causation. For decades, the dominant view of agent causation in action theory has been that it is reducible to event causation. In this regard, Velleman declares that ‘the proper goal for the philosophy of action is to earn the right to make jokes about primitive agent-causation, by explaining how an agent’s causal role supervenes on the causal network of events and states’ (2000, p. 132). I am skeptical that anyone has so far earned this right, and I argue that no postulated event or state in contemporary variants of event-causal theory seems sufficient to account for human agency. My sympathies lie with primitive agent causation, which faces, amongst other main challenges, the denial of an independently persisting subject of consciousness from Hume’s bundle theory of the self. I argue that this theory crucially lies on a mistaken phenomenology of consciousness (2.2).

2.1. Perhaps one of the best ways of illuminating the concept of human self-governance is by means of these three experiential gaps that a full and adequate account of human agency must fill in. In *Rationality in Action*, Searle argues that there are three gaps that are presupposed by normal cases of rational agency. First, in a deliberative or decision-making episode, there is a gap between your beliefs, desires and other motives and the actual decision that you make, where a decision consists in the formation of a prior intention. Second, once you have made up your mind to do something by forming a prior intention, there is a gap between this intention and its execution in action. And third, whenever you are in the course of some extended pattern of activity, there is a gap between its initiation and its continuation to completion (2001, p. 63).

These gaps - which I call 'gaps of rational and volitional activity,' because they are spaces of activity for reason and the will - do not seem to be creatures of fantasy. Their existence is attested by a wide range of phenomena, notably varieties of the pathology known as aboulia, which is one of the common disorders of diminished motivation. Consider, first, the first gap, or the gap between the motives or reasons for the decision and the decision itself. The phenomena of interest here are these two forms of aboulia.

First, there is aboulia as a cognitive deficit or an inability to generate options. We ordinarily choose or decide from options suggested to us by our preferences in conjunction with the situation we find ourselves in: for example, whether to eat dinner now or an hour later, and whether with chocolate fondant or with white chocolate cheesecake for dessert. But sometimes, in the absence of situation-generated options, we have to generate ourselves the options or construct the possibilities from which to choose. And this requires thinking or cognitive activity on our part. Certain aboulia people, who suffer from a paralysis of intellectual activity (thinking and imagination), cannot choose or decide what to do because of a failure to generate the options from which to choose. This is a form of pre-deliberative weakness (of reason or the will or perhaps both). Second, there is aboulia as indecision, irresolution or inability to decide. In this case, the individual clearly recognises and identifies the possible options generated by himself or proposed by the situation, but fails to choose from them. It could be because the agent judges the options or possible actions to be equally valid or on a par (as in the famous Buridan's ass case), or because of the fear of commitment, or because of some other internal factor constraining the will. This is a form of deliberative weakness (Proust, 2019, pp. 259-60). In both cases, the gap of

activity between the motives and the intention, decision or choice remains unfilled because the aboullic agent is *passive* or cannot *do* what is required to bridge the gap.

In the normal case, the human agent ‘does’ something or stands in a certain active relation to his desires or motives, a fact largely recognised by philosophers. For example, Setiya speaks of ‘the *active* and *reflective* character of the attitude we take to [our] reasons’: we recognise our reasons as such and we perceive them as being ‘in some sense ‘up to us’ - we decide *why* to do something’ (2007, p. 39). In a similar vein, Velleman notes that we ‘can reflect on [our] desires as grounds for acting rather than causes - as advisory rather than coercive -, and can then decide whether to act on them’ (1999, p. 4). Or that we do not ‘find that [our] decisions are straightforwardly dictated by the relative strengths of [our] desires’ (p. 178). Similarly, Searle observes that we normally sense alternative future decisions as causally open to us, and the reasons preceding them as not setting causally sufficient conditions for them (2001, p. 62).

So the gap between motives and intentions and the characteristic activity that fills it are essential features of our agency and self-governance. Arguably, it is because of its failure to recognise the existence of this gap as a space for the exercise of our rational autonomy that the standard (or desire-belief) theory of action faces the problem of *disappearing agency* or the problem of leaving the agent out of his actions. On this theory, the agent appears to be constrained or led around by the nose by his motives; his reasons and actions are not in any way open or up to him.

The second gap, or the gap between the intention to act and the actual initiation of the action, is also indicated by interesting phenomena. One of them is a third form of abulia - abulia as inability to translate decisions into action. An individual with this condition makes decisions, or forms prior intentions, but is unable to execute them, though there’s no external factor impeding this execution (Proust, 2019, p. 260). Impeding factors are typically internal. Proust (2004) reports that psychiatrists and neurophysiologists have observed that patients with frontal lesions often have a diminished or broken willpower while having a fully functional capacity to move, to think and to desire. In most cases, the patients experience a ‘lack of impulsion’, that is, they fail to initiate the execution of actions they have judged desirable. This problem is also widespread among autistic children, chronically depressed people and heavy consumers of opium and cannabis (pp. 128-9). But, presumably, not all sorts of weakness of will in this gap

have frontal or other lesions as their culprits. Think of such more familiar cases as procrastination and the weakness that consists in revising one's intentions in the face of temptations, both of which belong here. The causes of these failures to do the things one has decided to do are probably less dramatic. Whatever the causes of the varieties of weaknesses we experience in this gap, the gap suggests the dynamic character of agency: that our intentions and similar action-explaining motivational states do not causally suffice to yield actions. They need to be executed by an agent.

Finally, the third gap or the gap between action initiation and continuation to completion. The relevant variety of abulia here is the one that consists in a quick abandonment of the action initiated. Here, the individual perceives the possible option, decides or forms the intention to pursue it, and starts pursuing it, only to abandon it soon after. This abandonment might be due to the agent realising a discrepancy between the cost of the action (in terms of effort needed) that he imagined or anticipated and the real cost that he experiences after starting the action. Our natural inclination towards minimal effort may kick in or the instability of the agent's own character may get in the way at this point. Or the abandonment might follow from the agent's evaluative assessment of his decision. The action that previously appeared good to him might now appear to be bad or likely to lead to unwanted consequences (Proust, 2019, pp. 260-261). In either case, the individual is disappointed and has a change of mind or heart during the execution of the action. Experiences like these are indications of the existence of a gap between action initiation and action completion.

2.2. Now can we plausibly make sense of this complex phenomenology of our agency in purely state or event-causal terms? If so, what states or events or combination thereof is human agency fully reducible to? The earliest event-causal theory of agency is perhaps Hobbes's account. Hobbes offers, in *Leviathan*, a very parsimonious psychology of human agency, one with just two kinds of passions or motions of the soul: appetites and aversions. An appetite is the motion of approaching an object, while an aversion is the motion of retiring from an object (Hobbes, 1998, I, 6, §2). When these passions are in conflict, and a deliberation ensues, the appetite or aversion that wins over its rivals and leads to action, or omission thereof, is the agent's will. (Hobbes, *ibid.*, §53). Clearly, there does not seem to be an agent or an acting self in this account of deliberation and agency. If anything, there is only a spectator, and an indifferent one at that.

To the Hobbesian picture of human agency, the standard (event-causal) theory of action, which claims that the paired motivational states of beliefs and desires cause and sufficiently explain our behaviours, adds only instrumental beliefs, and in so doing it remains essentially Hobbesian, with a disappearing agent.

Later event-causal additions to the psychology of the agent mainly build on Frankfurt's (1988) basic intuition that a person is essentially characterised by a concern about his motivational states and by the reflective ability to take his own stand of endorsement or rejection of them. For Frankfurt, that by which a person puts his weight behind his desires, or withdraws it from them, is a distinctive higher-order attitude, which he identifies as a second-order volition (basically a second-order desire about first-order desires). For Velleman, it is a motive that is functionally identical with the rational agent, one that he cannot be alienated from while remaining a rational agent: the desire for self-understanding or to act for reasons. It is this desire 'that performs the characteristic functions of agency' (2000, p. 141), and produces our actions or our bodily movements in our name. I call these views Neo-Hobbesian.

What these Neo-Hobbesian views do is, roughly, akin to transforming the indifferent spectator on the Hobbesian picture into an interested or committed spectator by endowing him with second-order desires about his (possibly conflicting) first-order appetites and aversions. The difference between the Hobbesian agent and the Neo-Hobbesian agent can be accounted for in terms of the difference between two kinds of spectators at a horse racing event. One is indifferent to the outcome of the racing because there is nothing at stake for him in it. The other cares because he has picked a side or bet on one of the horses (thereby making it his 'agent' or 'representative' in the race -), and he is financially and emotionally involved. But both are spectators regardless: should this horse come out victorious, its victory is no more attributable to the better (as his action) than to the non-better. The only meaningful difference between the two is that the former might enjoy it more than the latter.

The executive and autonomous control that we exercise as we successfully navigate the three gaps of rational and volitional activity is not captured here. Consider, for instance, just the sort of executive control that we consciously experience in the space of the second gap when we move our limbs to execute our intentions. How can an event-causal theory explain this basic but

fundamental feature of our agency? Is it that our motivational states, or rather the events of acquiring them, ‘first cause certain [neurones] in the motor cortex of my brain to discharge, [then] these neural impulses are transmitted through the network of neural [fibres] all the way down to the peripheral control systems, which [finally] cause the appropriate muscles to contract,’ as Kim (2011, pp. 4-5) speculates? Or is it maybe that our motivational structure behaves more like a compressed spring of energy, and that actions result from this structure when the energy in it ‘is released in the appropriate direction by a suitable eliciting event, such as a thought or decision, or a perception of an opportunity to get what one wants, and guided in that direction by (above all) the agent’s beliefs’ (Audi, 1993, pp. 101-2)? Whatever the explanation, the claim that our intentional states cause or produce our bodily movements (in our name) sounds more like a denial of agentive executive control. And it’s not obvious that this is the right way to earn the right to make jokes about primitive agent-causation.

To fully and adequately make sense of the gaps of rational and volitional activity and our conscious experiences therein, we need, I think, the notion of self or agent as an irreducible psychological subject of reflexive consciousness, perception, belief, desire, memory, reasoning and willing, and as the irreducible entity operating in those gaps. I will not argue for, or provide an account of, agent causation here. I will address just one of the most important challenges it faces, namely, Hume’s flat and vigorous denial of the existence of such an entity. In *A Treatise of Human Nature*, Hume writes: ‘when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe anything but the perception.’ In fine, we (or, equivalently, our minds) ‘are nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perceptual flux and movement’. And these perceptions, moreover, ‘have no need of anything to support their existence’ (I, 1, 6). Clearly, Hume denies both the notion of the unity of conscious experiences and the notion that our conscious experiences have a single common subject (- the ‘self’).

My concern here is mainly the denial of a single common subject of consciousness. I think that Hume’s account of human consciousness is mistaken. His statement that ‘I never catch myself at any time without a perception, and never can observe anything but the perception’ (ibid.) is

partly true and partly false. The first part of it - 'I never catch myself at any time without a perception' - is true if interpreted as meaning that I can never be aware of myself by another means than a perception. Or that it is only through a perception that I am aware of myself. Against this interpretive background, the other part - '[I] can never observe anything but the perception' - will be false because my perception (of something) embodies an awareness of myself.

If my perception of something embodies an awareness of myself, then consciousness operates in two modes. In one mode, consciousness is awareness of object and their felt qualities through mental operations (such as seeing, hearing, touching, smelling, or tasting). In the other, consciousness is awareness of oneself as the subject of the awareness of objects and their properties. And these modes are concomitant; they cannot be separated. This is what Lonergan's analysis of human consciousness proposes². As he notes, in the very act of my perceiving an object I become aware not only of the object but also of the one (myself) perceiving and the act of perceiving (2002, p. 159). To illustrate, consider a conscious experience - say, yours - of looking at a glass of water on the table in front of you. What's going on in this experience of yours? Well, there are two main things going on as far as human consciousness is concerned. One is that by performing a certain mental act - namely, looking -, you have an awareness of a glass of water on the table. This mode of consciousness is commonly called 'intentionality' (or the mind's directedness toward objects). The other thing going on is that by performing that same mental operation (by which you are aware of a glass of water on the table) you are simultaneously aware of yourself and of attending to that object. This awareness of yourself is the experience of yourself as a subject of the awareness of a glass of water on the table. We may call this mode of consciousness 'self-consciousness'.

So human consciousness is both intentionality and self-consciousness. It is because of this bimodal character of the human consciousness that, 'in a given conscious experience, the subject can be conscious, as attending, and yet give his whole attention to the object as attended to' (Lonergan, 2017 [1971], p. 12). And it also makes it possible for us to objectify our experiences, make sense of them and develop reflective responses toward them and their objects. In fine, *pace* Hume, consciousness requires a single common subject because conscious

² I discuss further Lonergan's account of consciousness in Chapter 5, Section 1.

experiences are not only experiences of something; they are experiences of something by someone. And the data of some of these experiences - for example, thoughts and judgements - are not given to us; they are products of a cognitive processing by - to borrow a phrase from computer science - some unified operating system: the self.

3

The investigation of the nature of the human agent and agent causation in the previous section puts us in a position to decipher the phenomenon of interest to us in action explanation. In the question of what makes it the case that the event of your arm going up is something that you do rather than something that happens to you, I will take it that the phenomenon of interest here is an intentional action. This is mainly because the notion of intentionality is directly tied to (moral) responsibility, and informing our practice of assigning responsibility, praise and blame is presumably the main reason why we care about understanding action. Further, in taking intentional action as the object of interest in action theory and this thesis, I will also take it that an intentional action is an action done for reasons (3.1); that, necessarily, when you do something intentionally, you intend to do it (3.2.); and that there is no principled distinction between intentional action and self-governed action (3.3). Because each of these positions is controversial in contemporary philosophy of action, I briefly defend them in turns in what follows.

3.1. To defend the view that intentional actions are actions done for reasons, which I believe is the traditional view, I will begin with a statement of the necessary conditions for intentional action which every theory of intentional agency must attend to, flesh out and unify. I think there are three distinctive marks of intentional action. First, intention or purpose. Doing something intentionally is associated with doing it for a purpose: with intending or wanting to do it either for its own sake or as a means to an end. In this sense, intentional action is a species of purposeful behaviour. Second, awareness or knowledge: doing something intentionally is normally contrasted with doing it accidentally, or by mistake, or in ignorance. It is associated with doing it consciously. Third, control: an intentional action is something that we 'do' rather than something that 'happens to' us. It is something in our control. From this it seems plausible

to infer that three features are required for doing something intentionally: a *purpose*, *knowledge* or *awareness* of what one is doing, and *control* over it³.

To see why we should accept purpose, knowledge or awareness and control as necessary ingredients of an intentional action, consider cases where one or two of them may be absent. Suppose, first, that you go to Morrisons to buy a crate of beer and as you're walking inside you suddenly don't remember what you want. This lapse of memory of the purpose of going to the supermarket may cause you to stop dead in your tracks. Here, you're aware that you're walking inside the supermarket and you're able to control your bodily movements, and yet because these movements are disconnected from their purpose, as a result of your lapse of memory, they become, in a sense, devoid of intentionality. Second, imagine that you tell the person standing beside you in the aisle of the bus that he is stepping on your foot. He instantly looks down and, confused, he replies sincerely: 'Sorry, I wasn't aware that I was stepping on your foot.' This confession of unawareness or ignorance would normally indicate to us that the person's stepping on your foot was not an intentional action. Third, let's consider the case of an individual with alien hand syndrome. (This neurological disorder deprives its victims of control over their anarchic hand.) And let's imagine an instance where the patient's hand uncontrollably shoots out to pick up food when he actually wants to eat. We can even imagine that there's some normal causal connection between the patient's desire and his hand's movement, and that the patient is aware of what his hand is doing. All this notwithstanding, his hand's movements must strike us as not being intentional simply because he cannot exert any volitional power over them.

Now the traditional view that intentional action is action done for reasons starts off from the assumption that human action is fundamentally and distinctively purposive. It is fundamentally purposive because we normally act for some purpose, with an end or aim in sight. The end of an action may be either what the action will bring about or only to perform the action. It is distinctively purposive because, unlike other instances of purposive movement - for example, '[t]he dilation of the pupils of a person's eyes when the light fades' (Frankfurt, 1988, p. 73) -,

³ Malle and Knobe have identified a higher number of the necessary conditions for intentional action. 'In people's folk concept of intentionality,' they write, 'performing an action intentionally requires the presence of five components: a desire for an outcome; beliefs about an action that leads to that outcome; an intention to perform the action; skill to perform the action; and awareness of fulfilling the intention while performing the action' (Malle and Knobe, 1997, p. 111). I think that there is some double counting involved here.

our bodily movements are under our guidance in a special way. Further, the traditional view construes the purposes that drive our various intentional actions as reasons for which these actions are performed. Taken together, these two propositions yield the thesis that intentional action is action done for reasons.

The thesis that intentional action is action done for reasons is espoused by many philosophers, including not only Anscombe (1963), Davidson [2001 (1963)] and Goldman (1970) but also Mele (1992b), Raz (1999) and O'Shaughnessy (2008). But it is disputed by many philosophers too. Dissenters contend that while all instances of acting for reasons are cases of acting intentionally, the converse is not true: some cases of intentional actions are not instances of actions for reasons. They hold that while all intentional actions that are not intrinsically motivated - that is, that are performed for a further end - are actions for reasons, those that are intrinsically motivated (that is, that are performed as ends in themselves) are not actions for reasons. Stock examples of such actions are cases of doing things simply out of an urge to do them, without seeing anything good about (doing) them. Such examples include idly kicking the leaves piled at the side of a path as you are walking along, or doodling just to pass time (Setiya, 2003, pp. 341-2; Heuer, 2014, p. 294). They also include cases of purely expressive actions, or actions that people perform while in the grip of some emotion or inclination: for example, drinking paint when you are seized by a sudden desire to do it (Wiland, 2012, p. 51; Hursthouse, 1991, p. 62; Alvarez 2009, pp. 300-1).

In defence of the traditional view, two points can be made. First, the objections to the rationality of intrinsically motivated actions are not problems for this view. They are no more than challenges to specific theories of reasons for action, namely desire-belief conceptions and conceptions of reasons as good-making features of actions. The challenge against desire-belief conceptions is that the explanation of intrinsically motivated actions cannot make reference to appropriate instrumental beliefs, or so argue some philosophers (Hursthouse, 1991, p. 59; Audi, 1986, pp. 541-4; Alvarez 2009, p. 299). In other words, there is no ground for the ascription of appropriate beliefs to their agents - for example, a belief about the value of what the agent is doing (Hursthouse, pp. 59-60) - which would make it possible to give these actions a standard belief-desire rationalising explanation. In response, this criticism, plausible as it is, does not show that the traditional view is false. What it shows is simply that not all reasons for action are

desire-belief pairs. And some proponents of belief-desire conceptions of reasons actually admit this much. They recognise that desires alone (Mele, 1992b), but also emotions and feelings, are often invoked as reasons, as in ‘I hit him because he insulted my wife and that made me angry,’ or ‘He jumped up and down for joy’ (Kim, 2011, p. 196). So there is no trouble here for the traditional view.

As to conceptions of reasons as good-making properties of actions, their problem is, according to dissenters, that expressive actions have no desirability characteristics; there is nothing to be said in favour of doing them (Alvarez, 2009, p. 298). This objection presupposes that reasons for action are normative reasons, that acting for reasons is acting under the guise of the good. In response to this, it should be noted that the traditional view in itself does not imply, or is not committed to, the guise-of-the good thesis. (According to one version of this thesis, when we act intentionally, or for a reason, we see that action as good.) It does not seem to follow from the claim that intentional action as action done for reasons is inherently purposive that all our purposes must be good or virtuous. But individual proponents of the traditional view take different stances on the relation between acting for reasons and the guise-of-the-good thesis. Some explicitly reject this thesis. This is the case of Mele (1992), who holds that ‘the reasons for which we act are not, even by our own lights, good ones. They might even be quite perverse. But perverse reasons are reasons nonetheless’ (p. 361). And others, for example Raz (1999), explicitly endorse and defend it. It is to defenders of the traditional view coupled with the guise-of-the-good thesis that the objection - that intrinsically motivated actions such as expressive actions are intentional actions but not actions for reasons - is addressed. These actions, it is claimed, are not normally ‘seen as good’ by their agents.

I am also a defender of the traditional view cum the guise-of-good thesis. This thesis, I think, is our best hope for meeting the explanatory or intelligibility constraint on reasons for action. According to this constraint, a reason for action must be the sort of thing that makes intelligible to the agent the action for which it is a reason. Now on the guise-of-the-good-thesis, for an action done for a reason to be intelligible to its agent just is for it to be seen as good by him. It is controversial amongst the proponents of this thesis what sort of mental state is this ‘seeing’ as good. For some, the state of ‘seeing’ something under the guise of the good is a desiderative state (a desire or a desire-like attitude). For others, the appearance of the good is a doxastic state

(a belief or a belief-like attitude). I think, however, that we should dismiss this dichotomy as being false and speak instead of two different but related guises of the good - the affective guise and the cognitive guise. In my view, the affective guise is the primary guise and the cognitive guise is the secondary guise, and the affective guise is the basis of the cognitive guise (about this more in Chapter 5, subsection 2.1).

If this construal of the guise(s)-of-the-good thesis is plausible, it may help address the objection that intrinsically motivated actions like expressive actions are not normally ‘seen as good’ by their agents. Consider, for instance, such purely expressive actions as these: in a wave of rage and violence toward Jill, Jules gouges out Jill’s eyes in her photograph; Tim jumps up and down upon hearing some good news; and similar ‘arational actions’ that Hursthouse mentions in her discussion (1991, p. 58). When critics of the traditional view claim that these actions are not actions for reasons because they are not ‘seen as good’ by their agents, they generally mean that these actions are not performed under the cognitive guise of the good. As Hursthouse notes, we cannot ascribe to the authors of such actions a belief about the value of what they are doing (p. 60). In defence of the traditional view cum the guise-of-the-good thesis, we may admit this rejection of cognitivist interpretations of the thesis,⁴ and argue that authors of purely expressive actions perform them exclusively under the affective guise of the good. As I argue in Chapter 5 (subsection 2.1), we have a direct perception of value in and through our feelings or emotions. And these affective states can independently motivate and rationalise actions. And this concludes the first remark in defence of the traditional view.

Now the second remark to make about the objection that intrinsically motivated actions are intentional actions but not actions done for reasons is this: the dissociation of intentional action and action done for reasons by critics of the traditional view results from a mistaken conception of intentional action, one which leaves out its purposive character. For example, Alvarez (2009) distinguishes two minimal (necessary) conditions for doing something intentionally: some awareness of what one is doing and some control over it (p. 295). And so does Heuer (2014), who claims that an agent acts intentionally if, and in so far as, he has belief-control over his behaviour (p. 298) or if ‘the agent has control over his bodily movements when he A’s and has a

⁴ For a defence of the ascription of evaluative beliefs to agents of purely expressive actions, see Raz, 2011, ‘On the Guise of the Good’ in *From Normativity to Responsibility*, pp. 59-84.

self-referential belief that he is A'ing' (pp. 296-7). On these views, intentional actions do not seem to be necessarily goal-oriented. The denial (or, perhaps, omission) of the purposiveness of intentional actions makes it hard to understand why these actions are characteristically under their agents' control. This is easy to understand on the view that intentional action is action for a reason (and vice versa) because human agency is essentially purposive. Agents control and guide their actions in light of their purposes for acting. Recognising this vindicates the view that an intentional action just is an action for reasons. When we act intentionally, we act with some goal in view and that goal is the reason for which we act.

3.2. The next position about intentional action that I want to defend here is that doing something intentionally necessarily involves intending to do it. Suppose that I now go to the kitchen to make myself a cup of tea. On this view of intentional action, I go to the kitchen intentionally only if I intend to go to there. Now to go to the kitchen, I need to get up from the desk and start for the door, open the door and walk down two flights of stairs, and then push open the kitchen door. Do I step on the bottom stair intentionally? And do I intentionally make the stairs creak as I walk down to the kitchen?

On the view that doing something intentionally requires intending to do it, it might seem that in going down to the kitchen I do not intentionally step on the bottom stair or make the stairs creak. This is because I do not have an intention (specifically) to step on the bottom stair or make the stairs creak. What I intend is to go to the kitchen. But this would be a too rigid interpretation of the view. On a non-rigid, more natural reading of this view, I intentionally step on the bottom stair, but I do not make the stairs creak intentionally. The reason for this difference is that the action of stepping on the bottom stair is a constituent component of my intentional action of going to the kitchen, whereas the action of making the stairs creak is only its side effect. My intention to go to the kitchen includes within its scope the act of stepping on the bottom stair but not the act of making the stairs creak. So on the view that doing something intentionally necessarily involves intending to do it, the intention in virtue of which the action is intentional extends to the component acts of that action.

I want to address here two sets of objections facing this view of intentional action. The first set concerns the issue of how intention relates to action in an intentional action. This relation has a

bearing on the assignment of responsibility for intentional action insofar as it is crucial for the distinction between what is done intentionally and what is not. Since Bratman brought this issue to the fore in *Intention, Plans, and Practical Reason* (1987), there has been a philosophical debate between two main views. On one view, there is a tight connection between intention and action: to A intentionally one needs to intend to A. This is the Simple View. On the other view, there is a loose connection between intention and action: to A intentionally one need not intend to A but only to have an A-relevant intention. Let us call this view the Complex View.

Bratman argues that the Simple View fails on three scores. First, it flouts the demand for strong consistency that applies to intentions because it entails the possibility of intending two contradictory states of affairs (p. 114). Call this *the consistency problem* for the Simple View. Second, there are cases of doing things intentionally without intending to do them: for example, side-effect actions or things that we do as a result of doing other things. The Simple View cannot account for the intentionality of such action-types. Call this challenge *the side-effect-action problem* for the Simple View. Third, intentions are characteristically associated with distinctive reasoning and motivational regularities. The Simple View undermines these behavioural regularities. Call this *the problem of intention-related regularities* for the Simple View (p. 125). I examine and refute these objections in sub-section 3.2 of Chapter Two.

The second set of objections facing the view that doing something intentionally involves intending to do it concerns its status as a causal theory of action, and, more specifically, an agent-causal theory. Causal theories advance, inter alia, that prior causal histories are what distinguish actions from mere happenings. The status of my view as a causal theory is clear from two claims made earlier: (i) The phenomenon of interest for action theory in Wittgenstein's question is an intentional action. And (ii) the complex phenomenology of human agency is most plausibly captured in agent-causal terms. Now Frankfurt (1988) has raised the following three concerns about causal theories. First, insofar as causal theories seek to explain human events in terms of what is extrinsic to them (causal antecedents), they are vulnerable to the objection from deviant causal chains: it is easy to imagine cases in which the effect of any postulated causal antecedents is something other than an action (p. 70)⁵. Second, the contrast between mere

⁵ In a discussion, James Lenman rejected this objection by remarking that certain things can be explained by factors external to them: for example, sunburns.

happenings and actions can be observed in the lives of spiders, so it 'cannot be explicated in terms of any of the distinctive higher faculties which characteristically come into play when a person acts' (p. 78). Third, the attitudinal conditions of a person's action may themselves be alien to him (p. 77). For Frankfurt, rather than any species of mental antecedents, the property of being under the person's guidance during its occurrence is 'the most salient differentiating characteristic of action' (p. 71).

I think that none of Frankfurt's objections poses a serious problem for the thesis that when you do something intentionally, you intend to do it. Regarding the problem of deviant causal chains, Setiya (2007) deflated it by arguing that deviant causation doesn't require rejecting the causal theory of action, but only the view that intentions or similar motivational states are triggers of action. A more plausible view is that these mental states trigger and guide action. Setiya pointed out that when you act (intentionally), the intention, purpose or goal does not just trigger the action; it guides it throughout. I want to buttress this point by remarking that if our intentional states - intentions, decisions, desires or volitions - were just mere triggers of action, they would not have the following characteristic motivational qualities: (a) they often increase in strength over time (until the pursued goal is attained); (b) they make us persist in our goal-pursuit in the face of obstacles; and (c) they make us resume our goal-directed actions after interruption (Chartrand & Bargh 2002, pp. 19-21). These three properties of motivational mental states show that they do much more than trigger actions; they also sustain and guide actions toward the achievement of desired outcomes.

Frankfurt's next objection to causal theories of action is that prior intentional states are irrelevant to the distinction between mere happenings and actions because spiders can and do act and yet they are not capable of such states. This objection can quickly be put to rest by noting that most of us, philosophers, do not seem to know much about the minds and behaviours of spiders to warrant the claims we are inclined to make about them. Our claims are mere speculations. However, scientists who empirically study spider behaviour have reported that spiders show a rich and extensive behavioural repertoire, great cognitive abilities (as evidenced by their capacity to adapt their behaviour to their situational needs), and a considerable variation in different behaviours (including foraging, web-building, courtship and mating) between members of the same species and between different species (Herberstein & Wignall, 2011, pp. 1-2). These

findings appear to deny any ground to the claim that spiders do not have the sort of mental faculties and states that feature in causal theories of action. If spiders have such capacities and states, it is reasonable to suppose that they play a crucial role in spiders' initiation, guidance and monitoring of their actions, and so must be relevant to the distinction between what spiders do and what merely happens to them.

Finally, let's consider the objection that the differentiating causal states or events can themselves be attitudes or motives from which the agent feels alienated, so that the agent might justifiably not see the behaviours he causes as his own actions. An illustration of this objection is the classic case of an unwilling addict: he craves a drug, tries to free himself from the craving, but ends up yielding to it. Frankfurt (1988) characterises him as 'a helpless bystander to the forces that move him' (p. 16) - to 'alien' forces or attitudes rather than his own. Now on Frankfurt's objection to causal theories of action, it seems that, though the causal history of the unwilling addict's taking the drug features his overpowering desire as one of the causal antecedents of his behaviour, this behaviour is not strictly an action of the addict insofar as he is alienated from its causal forces.

Two lines of argument can be presented against Frankfurt's objection from the possibility of alienation. The first is that no matter how alienation is plausibly construed, it is immaterial to the issue of the status of a behaviour as a mere happening or an action and to the general issue of intentionality. That is, whether we are alienated or not from the mental states that are causally related to our behaviours has no bearing on the character of these behaviours as actions. It is no part of the considerations in virtue of which events are (not) actions. A second line of argument is that, *pace* Frankfurt, there is a sense in which the unwilling addict is not a mere bystander to his akratic action. He is meaningfully in control of it. In order to shoot up heroin, the addict must execute a goal-oriented sequence of operations which he mentally has access to. This involves performing each constitutive operation in the right order and in the right way, and guiding and monitoring the movements of his arm at every step of the process to ensure a successful completion of the whole sequence (Compare Schroeter, 2004, p. 641). And this is how the drug addict exercises meaningful control over his behaviour. I will come back to this case shortly (in sub-section 3.3).

I will conclude my discussion of Frankfurt's objections with a brief remark on his contention that the property of being under the person's guidance during its occurrence is 'the most salient differentiating characteristic of action' (1988, p. 71). As I have already observed, control or guidance is one of the necessary ingredients of an intentional action. I think that, by making control the most salient and distinguishing feature of an intentional action and by disregarding mental antecedents, Frankfurt cannot account, in a non-circular fashion, for the intelligibility of action. Part of accounting for the intelligibility of an action involves explaining why an agent exercised his capacity for control over his bodily movements rather than behaving otherwise. Control cannot be the answer here. Only purpose or intention can. It is that in light of which control is exercised in the first place. Control presupposes a purpose: control is exercised in light of and for the sake of a certain purpose. Purpose shapes and guides the exercise of control. If so, control goes hand in hand with purpose and is subordinate to it. As such, it cannot be 'the most salient differentiating characteristic of action'.

3.3. Now the last position about intentional action that I want to defend here is that it is not, in principle, distinguishable from the so-called 'autonomous' or 'self-governed action'. There is a widespread agreement that there are two distinct, but perhaps related, projects in action theory: one is to explain intentional action, and the other to explain self-governed action. The distinction between these two categories is generally taken to originate in, and be supported by, Frankfurt (1988)'s case of an unwilling drug addict, which I briefly discussed earlier. In this connection, it is widely held that cases of failure of control over one's motivational states suggest a distinction between merely motivated action and full-blown action (Bratman, 2006, pp. 91-2), between intentional behaviour and free action (Watson 2004, p. 13), between 'run-of-the-mill intentional action' and 'full-blooded action' (Aguilar et.al., pp. 2011: 5), or between 'mere actions' and 'full-fledged actions' (Katsafanas, 2013, p. 121). Setiya (2007) acknowledges this distinction when he writes: 'my topic is intentional action, as such, not "autonomous action"' (p. 61). In light of this general recognition of the distinction, Stout (2005) is mistaken when he writes that 'intentional action' and 'full-blown action' are normally taken to be the same (p. 3).

I think that there's some confusion involved in this distinction, and I will make two points to clear it up. First, the distinction is predicated on the perceived inadequacies of Davidson's causal theory of action. The lesson that seems to have been drawn from them is that this theory

accounts for defective actions rather than human actions par excellence (Velleman, 2000, p. 124) or that it successfully explains intentional action but not autonomous action (Davis, 2010, p. 34). But this lesson is misconceived. The standard theory fails to capture human agency, pure and simple - that is, the agency of the sort of creatures that we are, creatures who are capable of self-governance. And this failure, I think, is due mainly to the theory's exclusion of the concept of the will. Davidson didn't want to invoke what he called 'mysterious acts of the will, or kinds of causation foreign to science' (2001, p. 83), unlike his philosophical predecessors.

As Price (2010), quoted at the opening of this chapter, notes, philosophers traditionally used the concept of the will to distinguish action from mere bodily movement (p. 29), and as the capacity, together with reason, thanks to which we are capable of self-governing agency. This conception of human beings and human agency was, of course, rejected by modern psychology, which derided particularly the notion of free will. A typical case of this rejection is embodied by behaviourism, which ignored the contribution of cognitive and volitional mental processes to human behaviour (Baumeister et.al., 2009, p. 487). The rejection also occurred in philosophy, especially following Ryle's attacks on the 'myth of volition'. But following the demise of behaviourism - it proved inadequate for the complexity of human behaviour -, there was a revival of traditional volitional concepts in psychology. In parallel, there has also been a revival of volitional concepts and of talk of free will in the philosophy of action -, a trend predicated on the limitations of Davidson's theory of action. In fact, Davidson himself later attempted to re-conceptualise intention, treating it as a distinctive pro-attitude, in order to secure 'the possibility of autonomous action in a world of causality' (2001, p. 88). It is very ironic that the failure of his earlier theory is seen as warranting the distinction between intentional action and autonomous action for human agents.

Second, if any distinction needs to be made here, it must be within the broad category of intentional action. It is undeniable that all self-governed (or self-controlled) actions are intentional actions. But, in a certain intuitive sense of self-governance, there are intentional actions that are not properly self-governed. Paradigmatic cases of such actions include those that are explained by such conditions as compulsion, mania and intoxication, which we usually perceive as impairing our capacity for self-regulation. It is presumably this intuitive sense that informs our social and legal practices of assigning responsibility. These practices factor into

their judgements not only the nature and effects of actions but also the mental states of their agents.

Much of the motivation for the distinction between intentional action and autonomous action comes from the perception that Frankfurt's drug addict is passive with respect to his behaviour or he is led around by the nose by his unendorsed desire, a motive not endowed with the authority to speak for him. This perception makes sense on an event-causal theory of action. But it is mistaken. As I argued earlier, the drug addict meaningfully participates in the production and guidance of his behaviour. Acting on an unendorsed desire does not amount to being a bystander to one's bodily movements. It is not desires that move and control our limbs; it is we ourselves that do.

Moreover, acting on such a desire, or succumbing to a temptation, does not always mean that one cannot do otherwise. In *Critique of Practical Reason*, Kant invites us to consider a man who says that he finds his lustful inclination uncontrollable every time he sees a brothel. Kant suggests that this man would realise that he can control his inclination if a gallows were to be erected outside the brothel for him to be hanged on it once he had satisfied his lust (see 5: 30). What is true about this lustful man, and about all of us in most situations, is not that we cannot resist our inclinations, but that we often find them awfully painful to resist. And so an agent that fails at self-control does not necessarily cease to be autonomous. That is presumably why many kinds of self-control failures are not generally seen to count as a necessary or sufficient condition of exculpation. Just because someone claims to have been overwhelmed or violated by his impulses is generally not enough to determine us to let him off the hook of blame. So self-control failures cannot justify the distinction between intentional action and autonomous action.

Chapter Two: On Human Agency as Planning Agency: Three Problems for Bratman's Planning Theory of Intention and Action

The problem with [the desire-belief account of intention and intentional action is that it is] more appropriate for an understanding of agents who are not also planning agents - dogs and cats, perhaps. Yet we see ourselves, and rightly so, as planning agents (Bratman, 1987, p. 8).

It is a truism that human beings are planning creatures. We think about and plan for the future every day. This is because we normally conceive of our lives as extending beyond the present, and project ourselves into the future. Our plans for the future are commonly called intentions, and intentions are crucial for (understanding) intentional action. If plans for the future are intentions, they are crucial for (understanding) intentional action. But are intentions really plans for the future? And are plans necessary for (understanding) intentional action? I do not think so.

But Bratman has made the phenomena of plans and planning the cornerstone of his account of intention, intentional action and practical rationality. In his influential *Intention, Plans, and Practical Reason* (1987), he argues that the standard (or desire-belief) theory of intentional action needs a substantial extension (in both its descriptive and normative aspects) to make room for the sort of agents that we are: planning agents, or agents who need plans for intra-personal and interpersonal coordination. The extension of the descriptive aspect consists in recognising intention as a distinctive mental state alongside desire and belief, and that of the normative aspect in allowing that intention has a direct relevance to the rationality of actions.

In this chapter, I critically examine only Bratman's account of intention as an addition to the psychological or descriptive aspect of the desire-belief model, or of what sort of mental state it is. About the nature of intention, Bratman's account propounds three central claims: (1) Intention is inextricably tied to plans and planning (p. 2). More precisely, intentions are the building blocks of coordinating plans. (2) Intention is a distinctive pro-attitude involving a volitional and reasoning-centred commitment to action: intentions 'are conduct-controlling pro-attitudes, ones that have a characteristic inertia, and ones that play a crucial role as inputs into and constraints on further practical reasoning' (p. 37). And (3) intentions essentially facilitate intra-personal and interpersonal coordination. This is Bratman's planning theory of intention, which he dubs a

‘modest theory of the will’. It yields an account of intentional action that loosens the connection between intention and action: a person intentionally ϕ s only when she acts with some intention, but it needn’t be the intention to ϕ .

I argue that Bratman’s planning theory of intention and action faces three problems. First, the thesis that intentions are intertwined with plans and planning and that they are elements of plans is afflicted by *the problem of non-informativeness*. If there is such a thing as a plan as a mental state, it is properly understood in terms of the mental state of having an intention. And so the thesis is uninformative and viciously circular (Section 1). Second, the second and third central claims suffer from *the problem of non-representativity*: some of our intentions - namely, proximal intentions or intentions concerning the immediate future - are not associated with the functions and dispositions that Bratman’s planning theory claims to be characteristic of intentions (Section 2). And third, insofar as this theory takes having an intention as necessarily involving having a disposition for further means-end reasoning, the theory will fail to account for the sense in which spontaneous actions can be intentional. This is *the problem of the intentionality of spontaneous actions*. I conclude with a brief examination of the implication of Bratman’s planning theory of intention for the relation between intention and action and the related issue of the assignment of (moral) responsibility (Section 3).

1

‘Our commonsense conception of intention,’ writes Bratman, ‘is inextricably tied to the phenomena of plans and planning’. Further, ‘an adequate theory of intention needs to take seriously these phenomena and their relation to intention’ (1987, p. 2). For it is by way of these phenomena that ‘we can get a good theoretical handle on intention’ (p. 87). Bratman’s own thesis about the relation between intention and plans is that intentions are typically elements of coordinating plans. Needless to say, the truth value of these claims depends in part on the meaning of ‘plan’ as used in them.

There are two main ordinary meanings of plan: the *Oxford Dictionary of English* distinguishes between (i) plan as ‘a detailed proposal for doing or achieving something’ and (ii) plan as ‘an

intention or decision about what one is going to do'. Let us call (i) plan as a set of executive steps or operations, and (ii) plan as intention or decision.

Both of these meanings are used in some psychological and philosophical conceptions of intention that relate it to the notion of plan. Three such conceptions - call them *planning theories of intention* (henceforth PTI) - deserve attention and discussion here: (PTI 1) An intention is an uncompleted part of a plan whose execution has already begun (Miller, Galanter and Pribram, 1960). (PTI 2) An intention is an activated plan to which an agent has committed himself (Kuhl, 1987). And (PTI 3) intentions are elements or building blocks of coordinating plans (Bratman, 1987). I will argue that all these three conceptions of intention in relation to plans suffer from the *problem of non-informativeness*. The reason for this is that to define the mental state of intention along the lines of (i) simply misses the point, and to define it along the lines of (ii) is not especially illuminating.

Let us begin this argument against planning theories of intention with an illustration of the two ordinary senses of 'plan' and how they are connected. Suppose that, as I am brushing my teeth in the morning and thinking about the day ahead and the things to do, I decide to make spaghetti carbonara for breakfast, to reply (after lunch) to the dozens of emails I have been neglecting, and to go to a jazz concert in the evening. In so doing, I have formed intentions concerning these three actions. My intentions to make spaghetti carbonara, to reply to emails, and to attend a concert are part of my plans for the day. So intentions are plans. Indeed, we normally speak of our intentions in terms of plans (and vice versa). Thus, if my housemate proposes that we have a video game night, I may decline the offer by telling him that I already have plans for the evening, namely, that I intend to go to a concert.

Now having formed these intentions, or settled on these plans, I may anticipate the sequence of actions required to carry them out. For my pasta carbonara, for example, I may anticipate the need to begin with boiling a large saucepan of water, then to add a teaspoon of salt and spaghetti when the water is boiling, and so on and so forth. And as for my plan or intention to attend the concert, I may feel that I will have to buy the ticket online after breakfast, take shower and change into proper clothes later in the afternoon, call an Uber at 5pm, and then head to the concert hall. Each of these sets of actions that I need to perform in order to execute my

intentions or plans for the day (making carbonara and going to a concert) is also a plan. It is a plan in the sense of an organised sequence of operations for achieving something. And as the component operations are actions in their own right, each could be an object of an intention. So my intention to make pasta carbonara involves an intention to boil a large saucepan of water, and in intending to attend the concert, I must intend to get the ticket.

This illustration of the relation between intention and plan still does not yield any deep, non-trivial insight into the nature of intention as a mental state. The most we might squeeze out of it is perhaps the notion that an intention to A involves some knowledge of how to (try to) A, but not necessarily having a plan for the execution of that intention. And my contention here is that none of the sketched planning theories of intention - that is, PTI 1, PTI 2 and PTI 3 - yields such an insight.

1.1. Let us first consider PTI 1, which cognitive psychologists Miller, Galanter and Pribram propose in *Plans and the Structure of Behavior* (1960). According to these authors, we ordinarily use the term 'intention' to refer to an uncompleted part of a Plan that is being executed (p. 61). By 'Plan,' a key and technical concept in their theory, they mean 'a *hierarchy* of instructions' on how to perform a certain action or 'any hierarchical process in the organism that can control the order in which a sequence of operations is to be performed' (p. 16). Plan is 'essentially the same as a program for a computer' (p. 16). It guides behaviour like a programme guides an electronic computer (p. 2). On this view of intention, when we say that someone has an intention to do something, we mean that 'he has begun the execution of a Plan and that this intended action is a part of it' (p. 61). So the statement 'I intend to meet Max when I arrive in London' means that I am already executing my plan for travelling to London and that meeting Max is a part of this plan. In short, on PTI 1, we speak of people's intentions only with reference to the unfinished parts of the particular plans that they are already executing (p. 62).

As it stands, PTI 1 is obviously wrong on many counts. I will mention and discuss only two. First, PTI 1's behaviourist bias - the implication that in order to have an intention one must be exhibiting a relevant observable behaviour - makes it utterly inconsistent with our ordinary understanding of the phenomenon of intention and our experience of having an intention. On this theory, I did not form, and hence have, the intentions (to make spaghetti carbonara, to reply

to emails, and to attend a concert) while I was brushing my teeth because I was not executing any corresponding plans. And this is simply not how we ordinarily understand and experience intentions. Consistent with our understanding and experience is the fact that an intention is what sets the execution of a plan in motion, that we typically form intentions prior to acting, and that we often have intentions that do not result in actions. Another fact is that we can now intend to A later while doing nothing now as a means or as a preliminary step to A-ing (Bratman, 1987, p. 4). So the statement ‘I intend to meet Max when I arrive in London’ need not mean or imply that I am already on my way to London. I might be lying in my bed and telling someone on the phone that I intend to see Max when I arrive in London next week.

Second, some of the actions that we may perform with an intention do not require executing a plan or performing a set of operations. This is the case with the so-called *basic actions* (about which more in Chapter 3, subsection 1.1) - for example, lifting a finger. Contrast these with most of the physical actions that we perform in our daily lives. These actions are non-basic in a double sense. They are non-basic in the sense that we perform them by performing some other (basic) actions. For example, I turn on the light in my room by flipping a switch. And they are non-basic in the sense that they involve an organised sequence of operations, each of which may be a non-basic action. For example, cooking spaghetti carbonara or giving a lecture. Now on PTI 1, only non-basic actions can involve intentions (insofar as their performance consists in an execution of plans), and not basic ones. And yet there is no natural or principled reason whatsoever why I cannot intentionally lift my finger. So PTI 1 is seriously flawed. It completely misses the point about the nature of intention. And so it suffers from the *problem of non-informativeness*.

1.2. The same problem afflicts PTI 2, another token of the conception of intention in terms of a plan as a set of executive steps, operations or acts. Proposed by German psychologist Julius Kuhl (1987), PTI 2 is the view that intention is an activated plan to which an agent has committed himself, where a plan is (like on PTI 1) an organised sequence of instructions (for instance, motor programmes or mental operations). Put otherwise, a plan becomes an intention when it is in an activated state and when the agent is committed to its execution (pp. 281-2). Kuhl does not elucidate the concepts of activated state and commitment. Is a plan in an activated state when it becomes the object of an agent’s privileged motive? Does an agent become

committed to the execution of a plan when he chooses or endorses it? It is not clear what Kuhl's take on the first question might be. As about the second question, he provides this cryptic statement: 'Commitment to the execution of a plan can be encoded as part of the plan before it is activated or it can be generated after its activation, as a result of a series of mental operations' (p. 281). Now Kuhl proposed this planning theory of intention as an improvement of Miller and his colleagues' theory, which I have discussed as PTI 1. To some extent this is true, and owing mainly to the element of commitment (even though it is unclear what it amounts to in mental terms). But, overall, I think that Kuhl's view is also vitiated by its reliance on his predecessors' notion of a plan.

By defining intention in terms of a commitment to an activated plan, Kuhl implies that intending to A requires having a plan of action, and hence knowledge of the specific actions to perform and the specific occasions on which to perform them. This cannot be true of all our intentions. In particular, it cannot be true of those intentions that are general or wide-ranging or that concern a remote future (see Lewin, 1951, pp. 99-100). Consider, for instance, such general intentions as my intention to talk my cousin out of his intention to enlist in the army, or John's quite quirky intention to avoid being hit by falling objects. When I form the intention to talk my cousin out of his plan to join the army, I need not know exactly how to go about fulfilling it - whether I will first take him out for lunch or for a drink, at what point during our conversation I will bring up the subject, how I will introduce it, what line of reasoning I will develop, etc. As the actions involved in my intention are quite indeterminate, I cannot be said to have a plan for talking my cousin out of his decision. But not having such a plan surely does not entail that I do (or can) not intend to persuade my cousin not to become a soldier. The same goes for John's general intention to avoid falling objects. Forming such an intention does not require having knowledge of the specific action(s) and specific occasion(s) for its (their) execution. And so having this intention need not involve having a plan, or else being committed to a plan, for its fulfilment.

Now consider also a wide-ranging intention like the intention to be a caring older brother to one's younger siblings, or an intention concerning a remote future like the intention to stop being a philanderer when one is married and has kids. No plan needs to be represented, imagined or devised in order for an individual to form an intention of these sorts. Owing to the particular character of these intentions, and to human agents' rather severely limited epistemic

access to the future, it seems natural to expect that the relevant actions to be executed and the relevant occasions for their execution will be left unspecified at the time of intention formation. And this indeterminacy does not undermine the character of these intentions as intentions.

So it is wrong to construe an intention as an activated plan to the execution of which an agent is committed, as much as it is to conceive of it as an unfinished part of an unfolding plan. If there is any grain of truth in PTI 1 and PTI 2, it is little more than that intending requires knowing. As Raz puts it, 'one can intend to ϕ only if one has some knowledge of what ϕ -ing is, and if, when ϕ -ing is not a basic action, one has some idea how to try to ϕ . Absent such knowledge one cannot form an intention to ϕ ' (2022, p. 66). And a plan as an organised sequence of operations is no part, or need not be part, of such knowledge.

1.3. But how about PTI 3, Bratman's view that intentions are the building blocks of plans? PTI 3 differs from the planning theories just examined in its conception of plans as distinctive mental states rather than as sets or sequences of executive operations. This difference notwithstanding, Bratman's conception of intention in relation to plans suffers from the same *problem of non-informativeness* as the other theories. By understanding a plan as (i) 'a detailed proposal for doing or achieving something' and defining intention in relation to a plan, these theories simply missed the point about the nature of intention. And by understanding plans as mental states - a manoeuvre that echoes the dictionary meaning of plan as 'an intention or decision about what one is going to do' - and defining intentions as constitutive elements of plans thus construed, Bratman's theory is not especially illuminating. It simply leaves the concept undefined. In order for intentions as mental states to be parts of plans as mental states, intentions and plans must be mental states with at least some distinct properties. But there do not appear to be any such differences between intentions and plans in our ordinary understanding and uses of these terms. Neither are there in Bratman's own characterisation of these states.

What Bratman says about intentions applies to plans and vice versa. In trying to distinguish intentions from plans, Bratman (1987, p. 29) first stresses the similarities between them: as mental states involving a distinctive commitment to action, plans are 'intentions writ large,' and share the same supposed properties of intentions: they tend to be stable, control behaviours, and provide inputs into further practical reasoning. Then Bratman attempts to differentiate intentions

and plans in the following way: he claims that, 'because of their increased complexity (as compared with relatively simple intentions),' plans have 'other properties,' namely, they are partial and they are hierarchical. To grasp these properties, recall the example, at the start of this section, where I decide, amongst other things, to attend a jazz concert in the evening. Bratman would describe this decision as a partial plan - one that still needs to be filled in with details about which jazz concert to go to, how to secure the ticket, how to get to the concert hall, etc. But this decision or partial plan is (colloquially and uninformatively) just an intention. If so, partiality cannot be a feature of plans but not of intentions.

The same applies to the property of being hierarchical or having a hierarchical structure. Bratman says that means-related plans are nested within ends-related plans, *and* more specific intentions are nested within more general intentions (p. 29). But the conjunction operator 'and' is misleading if it is meant to suggest a distinction between plans and intentions with respect to structure. For the statement 'my intention to attend a jazz concert this evening embeds my intention to call an Uber at 5pm' is exactly equivalent to the statement 'my plan to attend a jazz concert this evening embeds my plan to call an Uber at 5pm'. If so, hierarchy cannot be a distinctive feature of plans either. Clearly, the properties of being specific or general, partial, hierarchical, and means-related or end-related attach in the same way to plans and intentions. Plans understood as mental states are thus one and the same thing as intentions. So to conceive of intentions as building blocks of plans is not to make an informative statement about them.

In light of the foregoing analysis, the first central claim of Bratman's planning theory of intention - that an adequate theory of intention must take seriously the phenomena of plans and planning (p. 2), or that it is by way of plans that we can get a good theoretical handle on intention (p. 87), and, relatedly, that our intentions concerning our future actions are typically elements in larger plans (p. 27) - appears to be completely uninformative about the nature of intention. For one thing, if there is such a thing as a plan as a mental state, it is properly understood in terms of the mental state of having an intention. And this precludes any talk of intentions as elements in larger plans, as the building blocks of plans or as the bread and butter of plans (if it purports to be informative). For another, not all intentions relate to, or involve, the phenomenon of planning. A simple, trivial intention like my intention to stick my tongue between my teeth in five seconds does not require any planning of how to carry it out. So we

need to look elsewhere, not at the phenomena of plans and planning, for informative insights into the nature of intention.

2

Now the other place to look for those insights in Bratman's planning theory of intention is his description of the commitment to action that is involved in an intention. Bratman characterises this commitment as comprising a volitional dimension and a reasoning-centred dimension, and sees these dimensions as being associated with specific functions or dispositions. The volitional dimension is associated with the role of intentions in controlling action. And the reasoning-centred dimension of commitment is associated with certain distinctive dispositions of the agent concerning his reasoning in the interim between intention formation and intention execution. These are the disposition to retain one's intention without reconsideration (except when there is a reason not to), and the disposition to reason from the retained intention to yet further intentions in a means-end deliberation. Bratman further claims that it is owing to this characteristic commitment to future action that intentions essentially facilitate coordination, both intra-personal and interpersonal (1987, pp. 16-7; pp. 107-9). I will take the claims about the commitment specific to an intention and about the coordination functions of intentions as the second and third central claims of Bratman's planning theory of intention.

I believe that Bratman is right that an intention involves a commitment to action in the sense in which a desire does not. An agent's desire to A merely predisposes him to A and does not settle for him the question of whether to A. By contrast, his intention to A commits him to A-ing and hence settles the question of whether to A. Many philosophers share this conception of intention as commitment (e.g. Harman, 1976; Velleman, 1999; McCann, 1991; Mele, 1992a), but it is unclear and controversial what the commitment characteristic of an intention actually involves or entails. I think that Bratman mischaracterises this commitment. In this section, I will argue that Bratman's claim about the functional and dispositional properties of the commitment involved in an intention also faces a problem: *the problem of non-representativity*. That is, the functions and dispositions that he identifies as characteristic of intentions are not embodied or exhibited by many of our intentions. I first discuss and dispute Velleman's construal of the volitional dimension (2.1), then examine the reasoning-centred dimension and the claim about

the functions of coordination (2.2), and finally address a potential objection to my arguments (2.3).

2.1. Let us start with the volitional dimension of the commitment characteristic of an intention. This dimension pertains to the putative fact that intentions are conduct-controlling pro-attitudes. Intentions control actions in at least two ways. The first way relates to the time of action. ‘If my future-directed intention manages to survive until the time of action,’ Bratman writes, ‘and I see that the time has arrived and nothing interferes, it will control my action then’ (1987, p. 16). The intention will control the action by motivating and guiding its performance in tandem with the relevant beliefs (p. 125). This intuitively plausible and fairly standard understanding of intention and action control is what we might call the guidance view of action control by intentions. On this view, when an agent A’s with the intention to do so, the intention guides the action, i.e. it keeps him on track towards a successful completion of the action (Raz, 2022, p. 69). Depending on the nature of the action, whether it is a simple or complex action, the guidance may consist in a feedback process that monitors basic movements from point X to point Y or the successful and orderly execution of the component acts of a complex operation.

Now if we go by this conception of intention and action guidance, it is not immediately obvious that action control is an exclusive property of intentions, nor that it is expressive of a commitment to action in any non-trivial sense. Arguably, many of our other motivational states can drive us toward action and control it quite in the manner described above. For example, a fleeting craving for the piece of cake on the table before me can drive me to pick and eat it while controlling my hand’s movements, and without me having formed a judgement or intention. Here, my action of picking and eating the piece of cake was caused and controlled, at least partly, by my craving, and so I cannot be said to have been committed to consuming it. If so, the capacity for action control and guidance does not buttress the claim that intentions involve a volitional commitment that no other pro-attitudes do. And this capacity cannot be what the kind of commitment involved in an intention actually consists in or amounts to.

The other way in which intentions control behaviours relates to the intervening time between intention-formation and intention-execution. Suppose that I now intend to travel to London next week. This intention will control my behaviour between now and the day and time of my travel

in at least two ways. First, it may give rise to distinctive dispositions (Bratman, 1987). One such disposition is the disposition to engage in a means-end reasoning concerning how to travel to London and how to get the tickets for my means of travel. And another is the disposition to retain my intention to travel to London, or to avoid forming intentions that are inconsistent with it, or even to endeavour to shield it from predictable temptations. I will be disposed to preserve my intention unless I have a reason to reconsider, modify or abandon it. Second, if it endures, my intention to travel to London will keep me alert to the time to go to the train or coach station, or to such eventualities as a cancellation or a delay, in order to avail myself of the means for its fulfilment. This is how intentions may be said to exert control on an agent's behaviours during the time leading up to their execution. As Bratman puts it, it is by way of the processes of reasoning and intention retention that intentions exert their influence on future conduct (p. 108).

This explanation of the volitional dimension of the commitment involved in an intention in terms of its conduct-controlling property in the interim between intention-acquisition and action performance appeals, mainly, to features of what Bratman identifies as the reasoning-centred dimension. It does not, however, seem plausible to conceive of deliberative dispositions as being essential to the commitment characteristic of an intention. As I will argue in the next subsection, there is a class of intentions that do not typically trigger any means-end reasoning about their realisation. In my view, the commitment involved in an intention is purely volitional in the sense that intending to A is being settled or decided upon A-ing. And my understanding of this commitment builds upon that of Korsgaard (2006). That is, following Korsgaard, but in a way detached from her Kantian ethical perspective, I hold that willing or intending to do something is committing oneself to do it in all the circumstances that are relevantly similar to the expected ones and unless there is a good reason why not (see Chapter 5, Section 3, and Chapter 6). But, now, I will assume the correctness of Bratman's conception of the volitional commitment of an intention just for the sake of argument - that is, in order to raise some worries for it.

I will raise just a couple of worries. The first concerns the relation between intention and belief in Bratman's account. It seems that if an intention is a pro-attitude with a volitional commitment to action it must involve a positive sort of belief toward the intended action. If my intention to travel to London next week will characteristically control my behaviours in the interim between my forming that intention and my executing it, it seems that the intention must involve the belief

that I will (try to) travel to London when the time comes. Without this sort of belief, it would be hard to explain why I am deliberating about the means to my end of going to London, why I am disposed to resist the temptation against going there, or why I will ensure that I do not miss my train or coach. Bratman (1987, p. 37), however, rejects coupling intention and belief this way. He claims that an agent's intention to A supports, but does not require, the belief that he will A. He thinks that one can intend to A and yet be agnostic about whether one will try to A when the time comes - as when a person consciously aware of his tendency toward absentmindedness intends to stop at the bookstore on his way home. He also holds that one can intend to A and yet be agnostic about whether one will succeed when one tries - as when one intends to carry out a difficult rescue operation. Bratman claims that these cases of intention in the face of agnosticism about the intended action do not involve any kind of criticisable irrationality. What is irrational is intending to A and believing that one will not A. And he thinks that these cases raise serious concerns for the view that an intention to A requires a belief that one will A.

These concerns, I think, are only apparently serious. The case of the rescuer does not seem to pose any threat to the association of intention and belief. Intending to do something does not require believing that one will succeed; it requires only believing that one will try. Arguably, believing that one will try to do what one intends is sufficiently consistent with the commitment involved in an intention. It can explain what I do in the interim between forming the intention to go to London and its execution. And believing that one will try is possible in the face of agnosticism about whether one will succeed. It seems, therefore, perfectly reasonable for one to intend to do something and be doubtful whether one will succeed (as long as one believes that one will try). A working-class and top-performing student from a State school in Sheffield who arrives at Oxford to start his PPE degree and feels out of place there can legitimately intend to get a First, and hence believe that he will try, while being in doubt about his success. So the possibility of intention in the face of skepticism about success is compatible with the coupling of intention with the relevant belief.

But intending to A while being doubtful that one will try to A when the time comes - as in the case of the absentminded agent - is problematic. It is problematic less for the view that intending to A involves believing that one will A - because, on this view, it can be argued that this agent merely wishes, rather than intends, to stop at the bookstore - than for Bratman's account. This

account construes the volitional commitment (involved in an intention) in dispositional and behavioural terms. The case of the forgetful agent raises questions about commitment and practical rationality for this account. If this agent truly intends to stop at the bookstore and is consciously aware that he might forget, and yet he does not do anything to ensure that he will execute his intention, in what way is he volitionally committed to stopping there? If he is volitionally committed to stopping at the bookstore, is he not guilty of a form of criticisable irrationality for not ensuring that his tendency toward absentmindedness does not thwart the execution of his intention? It seems hard to attribute to the absentminded agent a Bratmanian intention with its associated conduct-controlling property or practical dispositions. And the reason for this is that this agent does not have the right kind of cognitive attitude toward his intended action.

If it is hard for an intention to express a special kind of commitment to action if it does not involve a positive cognitive attitude toward the intended action, it is also hard for it to facilitate coordination without such an attitude. One of the central claims of Bratman's planning theory of intention is that intentions normally facilitate intra-personal and interpersonal coordination, and that they do so 'in part by providing support for expectations that they will be successfully executed' (Bratman, 1987, p. 37). My contention is that it is hard for intentions to support this sort of expectations and support coordination if they do not embody a positive cognitive attitude toward their objects. To see this, suppose that I tell you that I intend to throw a party for my birthday this weekend and that I do not believe that I will do it. On Bratman's view, this intention will provide support for your expectation that I will throw the party and inform and justify your decision to come to my party. But this does not seem to be how most people would normally react to my intention coupled with my skepticism about its object. Most people would probably take it that I am not yet committed to throwing a party for my birthday and so would not form any expectation or decision about it. If so, Bratman's account of intention as involving a commitment to action and playing a coordination function requires a positive cognitive attitude.

The second worry for Bratman's construal of the volitional commitment of an intention in dispositional and behavioural terms is that some facts about our experiences of intentions seem to challenge the characterisation of intentions as conduct-controlling pro-attitudes. We often

forget our intentions altogether (as in the case of Bratman's absentminded agent); we often fail to execute them (without any change of mind); and when we execute them, we often fail to carry it through or we often carry it through but only by forcing ourselves 'against powerful and persistent natural inclinations' (Frankfurt, 1988, p. 85). These are cases of the failure of intentions to control behaviours and lead to their execution. As such, they seem to cast doubt on the claim that intentions are conduct-controlling in the interim between their acquisition and execution or to suggest, at least, that intentions are not the source of their conduct-controlling power. If it turns out that intentions are not themselves the source of their conduct-controlling power, it would follow that they are not themselves conduct-controlling. And if intentions are not themselves conduct-controlling, then they do not involve a volitional commitment. In what follows, I will argue that intentions are not the source of their controlling power. So the challenge for Bratman is to dispute this claim or to accept it and yet defend his construal of the volitional commitment involved in an intention.

Now my claim is that whether intentions control behaviours depends more on factors external to them than on the nature or merits of the intentions themselves. More precisely, the relative stability and regulatory power of intentions over time are a function of two main factors. One such factor is the individual's character, habits and personality. These are normally more stable and better predictors of future behaviours than intentions. Experimental evidence from social psychology suggests that the intentions of people whose behaviour is more sensitive to internal states or dispositions than to situational cues or contingencies are more stable and more likely to lead to their execution than the intentions of those whose behaviour is more responsive to external cues (Ajzen, 1985, p. 22).

Another factor is the degree of commitment with which intentions are held. Intentions tend to be held with varying degrees of confidence. This should not sound odd to philosophical ears. It is now fashionable to see intention as a degreed state much like a belief. As a credence is normally taken to represent a degree of belief, or a level of confidence, in a proposition's truth, so is 'the commitment involved in an intention as a pro-attitude toward an action [...] taken to admit of degrees of confidence' (Jian, 2023, p. 362). A natural way to explain the degrees of confidence with which one holds one's intentions is to speak of how much one values, or cares about, the objects of one's intentions. On this view, to hold that the relative stability and regulatory power

of intentions depend on the degree of commitment with which intentions are held is to say that they depend on the strength of the needs or reasons that give rise to them.

The claim that the stability and controlling power of intentions depends on the underlying needs or reasons is supported by evidence from experimental studies of motivation for resumption in cases of interrupted intentional activities and cases of forgetting of intentions. This evidence suggests that the strength of the urge to resume the said activities depends directly more on the subjects' internal attitudes towards those activities (Lewin, 1951, p. 104) - that is, on the reasons or needs for which the subjects intended to perform those activities in the first place. The same evidence also suggests that both the forgetting of our intentions and our ability to carry them out in the face of obstacles depend on the strength of the needs underlying them (p. 112). This makes intuitive sense. Intentions are the sort of attitudes that are formed and held for reasons. Their temporal persistence and regulatory power should depend on the vitality of the reasons from which they arise.

Let me take stock here. In this subsection, I have examined Bratman's conception of the volitional dimension of the commitment involved in an intention. On this conception, intentions are conduct-controlling pro-attitudes: they control behaviours at the time of action and in the interim between their acquisition and their execution. At the time of action, intentions motivate and guide the intended actions toward their successful completion. And in the time leading up to their execution, intentions control the agent's conduct by giving rise to dispositions for means-end reasoning and for their retention and realisation. I have argued that motivating and guiding action to its completion is not distinctive of intentions; it is a property common to other motivational states. And, assuming for the sake of argument that intentions control behaviours by triggering distinctive dispositions, I have raised two serious worries for this construal of volitional commitment. First, Bratman's view that an intention to A does not require the belief that one will A makes it hard for an intention to express a volitional commitment to action and to facilitate coordination. Second, some of our experiences of intentions suggest that intentions are not themselves conduct-controlling, and so pose a threat to Bratman's construal of the volitional dimension of the commitment involved in an intention.

2.2. I'll now turn to claims about the reasoning-centred dimension of the commitment associated with an intention and about the functions of intra-personal and interpersonal coordination. Bratman holds that the reasoning-centred dimension involves means-end reasoning and intention retention dispositions, and that it is these dispositions that are critical to our understanding of intention (1987, p. 9). He also claims that intentions characteristically play in our lives the crucial role of promoting coordination, both intra-personal and interpersonal. As temporarily extended and social agents, we need to coordinate our present and future activities and to coordinate our activities with those of others in order to achieve our complex goals (pp. 2-3). Intentions help meet these needs 'in part by providing support for expectations that they will be successfully executed' and they provide such support because they are 'conduct-controlling pro-attitudes' that have 'characteristic inertia' (p. 37).

I want to argue that Bratman's claims about the reasoning-centred dispositions and coordination functions associated with intentions suffer from *the problem of non-representativity*. That is, they are not true of all cases of intentions. There are many instances of intentions, notably the so-called proximal intentions (about which more in 2.3), that do not play coordination functions or trigger means-end reasoning. First of all, we typically form intentions in situations in which we are deciding whether or not to do something. Suppose that you are unexpectedly presented with a plate of cookies and that instead of your hand spontaneously shooting out to grab a cookie, you find yourself trying to make up your mind whether to have one. Here, by making up your mind, you've formed an intention (to take or not a cookie). But this intention might not serve to coordinate any larger plans or to trigger any (further) instrumental practical reasoning (Velleman, 2007b, pp. 197-198). Relatedly, the same holds for those intentions that are expressive of our choices between two or more alternatives available to us. 'Eat in or take away?'; 'eat in, thanks,' I respond to the waiter as I form the intention to have my lunch inside MacDonald's, and then go and wait for it to be served. These intentions do not necessarily involve deliberative dispositions or play coordination functions.

Second, some of our intentions do not require further means-end reasoning because the intended actions are more routine: for example, when I decide to do my laundry, I anticipate the need to gather my clothes, put them in a laundry bag, walk to the laundry room, etc. But none of these parts of the execution of my intention requires further means-end reasoning because they are all

habitual activities. And it would be far-fetched to claim that instrumental practical reasoning necessarily followed every episode of my forming an intention to do my laundry before this behaviour became routine. There seems to be nothing odd about forming an intention about a complex action and executing it without the need to engage in reasoning about how to realise the intention. So the claims about the reasoning-centred dimension of commitment and the coordination functions are false about some of our intentions.

2.3. Now Bratman might object to my arguments that many of our intentions do not involve a reasoning-centred commitment as an essential component or a coordination role as a necessary function. He might do so by claiming that his theory is about the central case of intentions - that is, *future-directed intentions* -, whereas most cases of intentions that I discuss are *present-directed intentions*. The point here would be that the said component and function are the necessary, distinctive features of future-directed intentions, but not of present-directed intentions or, as Bratman puts it, 'intentions concerning what to do beginning now' (p. 4). But is this a valid distinction? What are we to make of the notion of an 'intention concerning what to do beginning now'?

It seems to me that Bratman's distinction between *future-directed intentions* and *present-directed intentions* could be understood in one of these two ways. One way it could be understood is in terms of Searle's (1983) distinction between *prior intentions* and *intentions-in-action* and Brand's (1984) distinction between *prospective intentions* and *immediate intentions*. The other way it could be understood is in terms of Mele's (1992a, 2019) distinction between *distal intentions* and *proximal intentions*. In Searle's and Brand's distinctions, intention-in-action and immediate intention are functionally identical with, and so can be redescribed as, what is traditionally called a 'volition' - a purely executive mental event. In his later work, Searle (1991) accepted a re-description of intention-in-action as volition, following an argument by O'Shaughnessy (1991). And Brand himself conceived of an immediate intention as the proximate mental cause of action, arguing that beliefs and desires cannot fulfil this function, that they figure instead in the non-proximal mental aetiology of action. Understood this way, the notions of immediate intention and intention-in-action are rather technical. If Bratman's own distinction is construed along these lines, then it cannot be used to rebut my charge of non-

representativity, because the relevant intentions that feature in this objection are not special executive devices but ordinary pro-attitudes.

The other option is the construal of Bratman's distinction in terms of Mele's distinction between distal intentions and proximal intentions. Both of these intentions are future-directed. The difference between them is that the former are for the non-immediate future, - for example, my intention to visit my family in the Congo in three months -, whereas the latter concern the immediate future, as when I intend to raise my arm now. As such, the immediate future of the proximal intention can be just a moment or a few seconds away. My intention to do something now is future-directed from the point of view of the temporal gap between the conscious experience of forming the intention and the experience of its fulfilment. Obviously, intention formation always temporally precedes intention execution. And it is always forward-looking. Now insofar as intending implies the future (which can be just a moment or a few seconds away), all intentions are, essentially, future-directed, and so it seems wrong to distinguish, as Bratman does, between future-directed intentions and present-directed ones. The temporal distance, or the amount of time that passes, between the formation of an intention and its execution seems irrelevant to the sort of thing an intention is or to how it works.

Another reason why making this distinction is wrong is that one and the same intention can have both a 'now' component and a 'later' component - that is, it can have both proximal and distal aspects. For example, I may have an intention to work on my thesis for four hours at a stretch, beginning now. Mele calls such an intention a 'mixed' intention (2019, p. 2855). Against this background, Bratman's claim - that 'If we are going to make sense of the notion of intending to do something, we must make sense of the notion of future-directed intention' (p. 4) -, understood as the claim that distal intention rather than proximal intention is the intention par excellence, is unjustified. And so his distinction cannot rebut the charge of non-representativity against the claims about reasoning-centred dispositions and coordination functions.

3

In the previous two sections, I examined the three foundational claims of Bratman's planning theory of intention. These claims are: (i) intentions are intertwined with plans and planning and

they are elements of plans; (ii) an intention characteristically involves a volitional and reasoning-centred commitment to action; and (iii) intentions essentially fulfil the function of facilitating intra-personal and interpersonal coordination. I argued that conceiving of intentions as building blocks of plans is uninformative about the nature of intention. It is akin to stating that intentions are nested within other intentions, which clearly leaves the phenomenon of interest unexplained and is not true of all intentions. And I argued that the claims about the reasoning-centred dimension of commitment and about coordination functions are not representative of the entire class of our intentions. In this last section, I explore the implications of Bratman's planning theory of intention for the nature of intentional action. In particular, I examine its implication for the intentionality of spontaneous actions (3.1) and also for the relation between intention and action and the related issue of the assignment of (moral) responsibility (3.2).

3.1. Consider such an action as spontaneously reaching up and catching a ball unexpectedly thrown to you. We perform many actions like this every day in our lives. And we perform them without conscious thinking or planning; instead, we seem to be driven by our situational impulses, or similar operating forces, and guided by our perceptions. Moreover, we typically do not experience our spontaneous actions as being unintentional, let alone as being mere happenings. Now if they are intentional, they pose a challenge to Bratman's planning theory of action. This theory holds that the mental state of intention is an essential component of the phenomenon of intentional action, and that it is intrinsically linked up with planning and involves distinctive deliberative dispositions. If so, the theory seems to exclude spontaneous actions from the domain of intentional actions. Or, at any rate, it may have difficulty explaining how spontaneous actions could be unplanned and yet involve intentions with their alleged associated dispositions and characteristic roles. Call this *the problem of the intentionality of spontaneous action*.

Bratman very briefly discusses this issue in *Intentions, Plans, and Practical Reason*, and proposes the following two solutions. First, many spontaneous actions are perhaps best described as being neither intentional nor unintentional, 'although they are under the agent's voluntary control and they are purposive' (1987, p. 126). Hence, they are not a problem for his planning theory because they are not intentional actions in the first place. They are simply out of

the scope of his theory. On this claim, your spontaneous action of reaching up and catching the ball might not be either intentional or unintentional. Second, in the cases where an agent acts intentionally by acting spontaneously, these spontaneous actions are intentional in virtue of the agent executing, perhaps, some long-standing general intention, one that applies to his present situation, rather than an intention formed in and for that situation (p. 126). So if in catching the ball spontaneously you are acting intentionally, it is more likely that you are executing a general intention (say, to protect yourself from objects flying in your direction) than that you are executing an immediate and situationally specific intention to catch that ball.

Bratman thinks that, jointly, these two solutions ‘go at least part of the way toward answering’ the objection from the intentionality of spontaneous actions, but he is ‘unsure whether such a defence can work for all cases’ (p. 126). But I do not think that they go even half way to saving his planning theory of intention and action from this trouble. Each of the points faces some objections. Regarding the first solution, by treating many of our spontaneous actions as purposive and agent-guided behaviours that are nevertheless neither intentional nor unintentional actions, it raises the question of what the necessary and/or sufficient ingredients of intentionality are. I argued in the previous chapter (sub-section 3.1) that purpose (or intention), knowledge or awareness and control are the necessary properties of an intentional action. On this view, only awareness of what one is doing needs to be added to the ingredients identified by Bratman. If this view is true, then many of our spontaneous actions are intentional, because we are typically aware of what we are doing when we act spontaneously. But if Bratman’s solution is plausible, that is, if many spontaneous actions are not intentional, then few of our routine actions are intentional. And so acting intentionally might seem to be an exception rather than a rule in our lives. This is an unacceptable consequence of Bratman’s first solution.

As regards the second solution, if it is acceptable, that is, if it is true that the spontaneous actions that are intentional owe this character to a long-standing intention, then they cannot be genuinely spontaneous. Many of our spontaneous and routine actions, notably everyday conversations, are not subserved by long-standing personal policies resulting from practical reasoning. That is at least part of what it means for an action to be ‘spontaneous’ (or ‘automatic’). Consider, for example, the case of this spontaneous action by Oscar Wilde that Arpaly mentions in her discussion of actions that we perform without deliberation. ‘When Wilde bragged that he can

make a pun on every subject, someone suggested the queen, whereupon Wilde responded very quickly - that is, with no time to deliberate [or even think] - by saying: “the queen is no subject” (2021, p. 189). There is no doubt that this action is both spontaneous and intentional, and yet it would not make any sense to attribute to Wilde any relevant long-standing general intention applicable to the situation of his (verbal) action.

The problem of the intentionality of automatic actions has led some philosophers to take the ‘mind out of action’ (Di Nucci, 2013) - or discount the agent’s mental states - and espouse Frankfurt’s view that the property of being under the agent’s guidance during its occurrence is ‘the most salient differentiating characteristic of action’ (Frankfurt, 1988, p. 71). As I argued in Chapter One (sub-section 3.2), this view is very implausible because control goes hand in hand with purpose or intention and is subordinate to it. This is required for the intelligibility of our intentional actions as actions done for reasons. I believe that the intentionality of our automatic or spontaneous actions is, at least partly, more plausibly explained by the habitual inclinations or willingness with which individuals antecedently are disposed to choose or pursue the objects of those actions. Once you are habitually inclined or willing to do something, you are normally led to do it without reflecting or forming an intention. Against this background, it is easy, I think, to understand how we act spontaneously yet intentionally, but without forming an intention to act or without our actions being preceded by a conscious intention.

It is true, as Frankfurt (2006) pointed out, that ‘rollicking spontaneity is not exactly the hallmark of our species’ (p. 1) because we, human beings, are self-conscious and self-reflective creatures. But it is also obviously true that we do not always engage in practical reasoning or consciously form intentions before we act. In particular, when we act spontaneously, we act without any prior intention or any intention associated with a distinctive deliberative disposition.

3.2. Finally, I want to examine the implication of Bratman’s planning theory of intention for the relationship between intention and action in the phenomenon of intentional action. This relation is important for the distinction between actions that are performed intentionally and those that are not. And this distinction informs our practice of assigning moral responsibility. There is a philosophical disagreement about how intention conceptually relates to action - a debate sparked and shaped by Bratman’s original treatment of the issue in *Intention, Plans, and Practical*

Reason. Bratman first identified a general conception of the issue - the Single Phenomenon View -, according to which, inter alia, it is the relation of an action to an intention that makes that action intentional. And then he distinguished two main positions within this conception. According to the Simple View, when someone A-s intentionally, he necessarily intends to A. According to (what I will call here) the Complex View, when one A-s intentionally, one need not intend to A; one nevertheless needs to have an A-relevant intention (Bratman, 1987, p. 112; Mele et.al., 2021, p. 761).

Further, Bratman argued that the Simple View fails on three scores. First, it flouts the demand for strong consistency that applies to intentions because it entails the possibility of intending two contradictory states of affairs (p. 114). Call this *the consistency problem* for the Simple View. Second, there are cases of doing things intentionally without intending to do them: for example, side-effect actions or things that we do as a result of doing other things. The Simple View cannot account for the intentionality of such action-types. Call this challenge *the side-effect-action problem* for the Simple View. Third, intentions are characteristically associated with distinctive reasoning and motivational regularities. The Simple View undermines these behavioural regularities. Call this *the problem of intention-related regularities* for the Simple View (p. 125).

In what follows, I want to defend the Simple View against these objections. I will not make a positive argument for the Simple View, but only show that they are implausible. Let me begin with the consistency problem for the Simple View. This is the objection that the rational requirement of consistency that applies to intentions is violated by the Simple View. Bratman illustrates this violation with the following case. Suppose I decide to play two video games in each of which I must guide a missile into a certain target. I decide to play them simultaneously because I am ambidextrous and I can play one game with each hand. Suppose, further, that the two games are related and structured this way: if one of the targets is hit, both games are over; and if both are about to be hit simultaneously, the machines just shut down, making it impossible to hit either target. As a player, I know all this and I know that it is difficult to hit either target. Now in light of this knowledge, I decide to give each game a try, that is, to try to hit each of the targets (target 1 and target 2). Suppose, finally, that I skilfully succeed in hitting target 1, so that it seems that I hit it intentionally. If so, on the Simple View, I must have intended to hit target 1. Given that I simultaneously tried to hit both targets, I must also have intended to hit target 2. But

given my knowledge of the rules of the games - in particular, of the fact that I cannot hit both targets -, my intentions are inconsistent, and so I am irrational for having them. Yet I am obviously not irrational because my strategy of giving each game a try seems perfectly reasonable. If I am not guilty of irrationality, then I do not have the intention to hit target 1 and the intention to hit target 2. So the Simple View is false. If not, it yields a form of criticisable irrationality (Bratman, 1987, pp. 113-5).

Now we can capture Bratman's argument against the Simple View in this video game case in the following reasoning:

- P1. On the Simple View, I must intend to hit target 1 and intend to hit target 2.
- P2. Having both intentions in the present situation is practically irrational.
- P3. I am actually not guilty of such irrationality: it is rational to give each game a try.
- P4. If I am not guilty of practical irrationality, I do not have both intentions.
- C. Therefore, the Simple View is false.

To defend the Simple View against the charge of yielding practical irrationality, it should be enough for a proponent of this view to show that P1 is false. And holding P1 to be false implies that, on the Simple View, the agent's hitting either target was not intentional.

Bratman argues that when I hit a target during the game, I hit it intentionally. My hitting it is intentional because, he says, I wanted it, it depended on my skills and trying, and it was guided by my perception. For him, uncertainty about success does not undermine the intentionality of my hitting the target. But I think we can plausibly argue here that all these ingredients do not guarantee that I will hit a target when I try or that it is really up to me to hit it when I launch a missile. I do not have the relevant degree of control over the trajectory of the missile, and hence the uncertainty about the outcome of my attempts. Because control is, as I argued in the previous chapter, a necessary component of intentional action, there is a certain intuitive and deep sense in which my hitting a target, when I do, is not intentional. When I say that I have hit a target intentionally, this is acceptable because I am speaking loosely. Actually, what I am doing intentionally in the relevant sense - one that involves control - is launching the missiles, aiming at the targets, or trying to hit the targets, but not hitting the targets. And so I cannot reasonably

be said to be intending to hit both targets. Because of it not being entirely up to me - up to my desire, my skills, my effort and my perception -, I can only intend to try to hit each target or to give each target a try. And that's why P1 is false. So the Simple View is not guilty of irrationality.

The notion of control can also enable us to address the side-effect-action problem, or the objection that the Simple View cannot account for the intentionality of the things that we do as a result of doing other things. To examine this problem, consider the following example by Bratman: I intend to run a marathon and believe that in running it I will wear down my sneakers. Presumably, wearing down my sneakers this way will not be intentional. Bratman suggests that I may well wear down my sneakers intentionally if, in addition to believing that in running the marathon I will wear them down, I am consciously aware of this while I am running and if wearing them down also independently matters to me. Actually, however, these additional states of my mind about the fact of wearing down my sneakers should not change what I intend in running the marathon. What I intend is to run the marathon. But, given my relevant beliefs and desires, I may, in running the marathon, intentionally wear down my sneakers without intending to do so (Bratman, 1987, p. 123). The Simple View cannot explain the intentionality of this side-effect action.

Now why should we think that my wearing down my sneakers in this case is an intentional action? Intuitively, my believing or knowing that I will wear down my sneakers by running a marathon does not make my actually wearing them down intentional. It is not intentional not only because it is an unwanted consequence of my running the marathon but also because it is something I have no meaningful control over. It is up to me to wear or not to wear my sneakers; but it is not up to me to use them for running a marathon without wearing them down. But Bratman thinks that if my mind is equipped with two further features - my conscious awareness (of my sneakers being worn down while I am running) and my seeing it under the guise of the good -, my wearing down my sneakers may be classified as intentional. But he does not tell us why these features matter or suffice for the intentionality of an action. Nor is it any clear from his theory of intention and action.

I think that the additional mental features - the conscious belief and the pro-attitude - do not make any difference for intentionality. To see this, suppose that a patient with alien hand syndrome can form a true belief about, or correctly predict, when his hand will anarchically make a certain movement (e.g. pick up the empty glass on the table). Suppose, also, that the patient likes to see his hand behave that way. He is quite puzzled and amused by the mysterious physiological mechanics underlying his autonomous hand movements. Imagine, finally, that the patient's hand moves and grabs the glass precisely at the predicted time, much to his enjoyment. It seems that the patient's having such a true belief and desire wouldn't suffice to make his hand movement an action, let alone an intentional action. The reason for this is, presumably, that the patient has no control over his hand. The hand did not move as a result of his trying to move it. It moved out of its own quirky volition. Now if relevant beliefs and desires fail to make a mere happening an action, it seems very implausible that they are sufficient to make an uncontrolled side effect an intentional action. We are inclined to believe that they will need to be paired with the agent's exercise of control over this effect in order for it to qualify as an intentional action.

Finally, let us look at the problem of intention-related regularities for the Simple View. Using the marathon runner case just examined, this objection to the Simple View can be presented in the following reasoning:

P1. The marathon runner intentionally wears down his sneakers.

P2. On the Simple View, A-ing intentionally involves intending to A.

C1. Therefore, on the Simple View, the marathon runner intends to wear down his sneakers.

P3. Intending to A involves being disposed to reason about the means to A and being motivated and guided by the intention at the time of A-ing.

C2. Therefore, on the Simple View, the marathon runner is disposed to reason about the means to wear down his sneakers and is motivated and guided by his intention to wear them down at the time of running.

P4. The marathon runner is not so disposed, nor is he so motivated and guided at the time of running.

C3. Therefore, the Simple View is false.

The reason for the falsehood of the Simple View is, for Bratman, that it ascribes to the marathon runner an intention which is not associated with its characteristic behavioural regularities (1987, p. 125). As it may seem obvious, Bratman's intention-related regularities objection builds upon, and presupposes the truth of, P1 - there are cases of doing things intentionally without intending to do them (e.g. wearing down one's sneakers in running a marathon) - and P3 (the volitional and reasoning-centred dimensions of the commitment involved in an intention). But I have just argued against P1 and I argued against P3 in the previous section. If my arguments are plausible, then the intention-related regularities charge against the Simple View fails.

Against this argument Bratman might contend that the rejection of the intentionality of side-effect actions seems inconsistent with our practice of assigning responsibility for intentional actions. He seems to think that our common-sense or folk-psychological scheme for classifying actions as intentional and assigning responsibility for them includes within the scope of intentionality 'all the consequences of an act that are foreseen as certain or probable' (Sidgwick, quoted in Bratman, 1987, p. 124). Against this background, 'it seems natural to classify as intentional [the marathon runner's] wearing down of [his] sneakers' (p. 124). Now I am skeptical that we ordinarily take side-effect actions as intentional actions, but this is an issue perhaps best left for experimental psychologists and philosophers to investigate.

Chapter Three: Agency as a Way of Knowing? Velleman on Action, Intention and Self-Knowledge

I think that the standard story [of human action] fails to include an agent - or, more precisely, fails to cast the agent in his proper role. In this story, reasons cause an intention, and an intention causes bodily movements, but nobody - that is, no person - does anything (Velleman, 2000, p. 123).

Our self-conception as human agents involves a commitment to the view of ourselves as self-knowing agents. We normally know (in some way and under some description) what we are doing intentionally. And it is necessary for an action to count as intentional that its agent perform it knowingly (or, at least, not perform it in ignorance). At first glance, this seems to be all there is to the relation between self-knowledge and human agency. But some philosophers think that the relation goes farther, or deeper, than this. Velleman, for example, has argued that in order to understand our agency we need to understand our self-knowledge ‘because autonomous agency is rooted in self-consciousness’ (1999, p. 5). He suspects ‘an intimate connection between an agent’s self-knowledge and the nature of agency itself’ (p. 18), and argues that agency is just ‘a special way of knowing’ (2007a, p. xiii). Call this view *epistemicism*¹ about human agency.

In this chapter, I examine Velleman’s epistemicism about agency through what we may take as its three constitutive theses: (i) Practical knowledge is ‘a very different knowledge’ (p. 18) - a kind of knowledge not acquired or justified in any of the ordinary ways. Call this thesis *exceptionalism about practical knowledge* (Section 1). (ii) Practical knowledge is embodied in the agent’s intention, which is a kind of belief or cognitive attitude. Call this view *cognitive intentionalism about practical knowledge* (Section 2). (iii) Acting autonomously is acting in accordance with one’s representation or cognitive awareness of one’s (strongest) motives (Section 3). I argue that both exceptionalism and cognitive intentionalism are implausible, and that epistemicism about agency does not secure any genuine self-determination for human agents.

¹ We should not confuse the term ‘epistemicism’ as used here with epistemicism as the theory of vagueness, according to which vague terms (for example, ‘tall’, ‘thin’ and ‘bald’) have sharply precise but unknowable boundaries. The two concepts of epistemicism are different and unrelated.

Let's begin our examination of *exceptionalism about practical knowledge* with an attempt to unravel the mystery of our supposedly special self-knowledge. The term 'self-knowledge' is standardly used by philosophers to refer to the kind of knowledge that we (sometimes) have of our own mental states - thoughts, feelings, beliefs, desires, intentions, etc. - and our actions. My focus in this chapter is mainly on the knowledge that we have of our intentional actions, which is also described as practical knowledge.

Now suppose that you are attending an auction at Auction House London. Suppose, further, that at some point you want to enter a bid for an antique alarm clock, and you believe that you can do so by raising your arm. And so you raise it. What, if anything, do you know about what you're doing here? One thing that we may say straight away is that you know that you are raising your arm during an auction at Auction House London. Or you know *what* you are doing. But, intuitively, this cannot be all there is to your knowledge of your action. Not only do you know that you are raising your arm during an auction at Auction House London, but also you will normally know *why* you are raising it - that you want to enter a bid for an antique alarm clock. In sum, we might say simply that when you act intentionally, you typically know *what* you're doing and *why* you are doing it. You know your intentional actions under this double description.

The next step in unravelling our mystery is to enquire about the nature of this knowledge we have of our actions. How can we plausibly characterise the knowledge we have of what we're doing and why we're doing it? Following Anscombe's treatment of practical knowledge in *Intention*, some philosophers argue that this knowledge has, or appears to have, 'certain philosophically interesting features, features that do not attach to one's knowledge' (Campbell, 2015, p. 3) of anything or much else. Let's call this view *exceptionalism about practical knowledge* (or *exceptionalism*, for short). Proponents of exceptionalism attribute the following three apparent features to practical knowledge: (i) practical knowledge displays first-person authority (1.1); (ii) practical knowledge is independent of perception or investigation (1.2); and (iii) practical knowledge is relatively immediate/spontaneous (1.3). I contest exceptionalism by arguing that (i) and (ii) are implausible, that only (iii) - psychological immediacy - is true, and that this property is not, however, distinctive of practical knowledge.

1.1. To begin with, does practical knowledge display first-person authority (FPA)? To say that practical knowledge has this property is to suggest ‘the possibility of an epistemic first-person third-person asymmetry with respect’ (O’Brien, 2007, p. 158) to an agent’s action. This asymmetry can be captured by the statement that, typically, the agent ‘X is in a better position than others to know [his action], because X knows [it] in a way in principle unavailable to others’ (O’Brien, p. 158). Stating this amounts to attributing to our knowledge of our (intentional or voluntary) actions a quality that is usually associated with our knowledge of our own mental and sensory states. About these states, the expression ‘first-person authority’ normally connotes that their owner enjoys a privileged epistemic access to them, that he is the best person to enquire from about them, that his sincere reports about them are presumed true, etc. (Campbell, 2015, p. 6). But do we really have first-person epistemic authority over the entire range of our intentional actions?

My contention is that the FPA thesis is either trivially true or merely implausible. It is trivially true about, for instance, our mental actions. Intentional actions can be mental or physical. Mental actions are of a purely psychological nature: they occur, well, in the mind - for example, searching for stuff in your memory or adding up numbers in your head. Intentional physical actions are typically embodied in bodily movements. Of these two classes of actions, the class of mental actions seems to be readily amenable to the FPA thesis because of the close resemblance that its members bear to the internal phenomena of intentional and sensory states. I normally know that I have got an occurrent desire to move my leg to kick the ball or to stretch it in a way that other people do not - that is, non-observationally and non-inferentially. The same goes for my mental actions: for example, searching for the name of someone from my distant past in my memory. We seem to be first-person epistemically authoritative with respect to the whole range of our mental states and actions.

But the FPA thesis is normally presented as a claim about physical actions, not about the whole range but only about basic physical actions. As O’Brien puts it, ‘[i]f there are descriptions of actions that are somehow basic, we *can* make our claim about our authority over our basic actions’ (p. 161). My contention is that, properly speaking, it is under this more modest guise that the epistemic asymmetry thesis is either trivially true or merely implausible. And this

depends on the kind of description under which we take our physical actions to be plausibly basic. Now there are roughly two potentially plausible theories of basic action in the philosophical literature. The first theory views basic actions as specific bodily movements. This view - call it *corporealism* - can be captured by the slogan: 'We never do more than move our bodies: the rest is up to nature' (Davidson, 2001, p. 59). And the second theory takes basic actions to be volitions, tryings or willings. This conception - call it *volitionism* - can be stated in the following slogan: 'We never do more than try or will: the rest is up to nature' (Ford, 2018, p. 702).

Now the FPA thesis is trivially true on a volitionist conception of basic actions. This is so because a volition, understood as the immediate and decisive causal antecedent of physical movements, is best construed as a mental event. And we are typically epistemically authoritative with regard to our mental actions under both the why-description and the what-description. So if volitionism is true, the FPA thesis is trivial. O'Brien, an exceptionalist about practical knowledge and a believer in the FPA, objects that volitions or tryings do not seem to be the description under which our actions are plausibly basic, that they do not constitute 'a narrow enough class of actions to be those over which we are authoritative' (2007, p. 163). Her point is that even if we are authoritative about our tryings, we need not be so about whether we succeeded in trying or, more precisely, whether our tryings produced their effects. You may know authoritatively that you tried to raise your arm to enter an auction, but this does not entail that you know that you raised your arm. True, but this is not at all an objection to volitionism. It is one to corporealism. To anticipate, I will argue shortly that the FPA thesis is implausible on a corporealist construal of basic actions, which is the one that O'Brien espouses. On a volitionist conception, we are authoritative over our tryings - the description under which our physical actions are basic -, but not over whether our tryings are physically or externally successful. So O'Brien's objection is misguided.

O'Brien also objects that, if tryings are personal-level mental precursors of actions, it is possible to deny that they are involved in every successful action production: in some cases, we can simply act without trying (p. 163). She has not provided any such cases, but we can draw some from Audi (1993), who holds the same view. Audi gives the following example. Suppose that you intend to raise your hand - say, to waive to an acquaintance who is passing. If you know that

you have the normal control of your limbs, or if you know that you can or will (effortlessly) raise your hand, you would surely not believe that you will try to raise your hand (p. 58). In response, it can be argued that, as an action-theoretic term of art, a trying needs not be understood as being essentially phenomenally effortful - a connotation that it has in our ordinary use of the word. In fact, other technical phrases taken to be equivalents of tryings or volitions - such as Brand's immediate intentions or Searle's intentions-in-action (see Subsection 2.3. in Chapter 2) but perhaps not O'Shaughnessy's 'bodily or physical will' (2004, p. 197) - do not have such a connotation. If this is true - if not all our tryings are effortful -, there is nothing odd about claiming that we try to raise our hand even when we justifiedly believe that our limbs are functionally normal or that raising our hand will not cost us any effort.

But if volitionism is actually false, then the FPA thesis is merely implausible on a corporealist construal of basic actions, or so I think. O'Brien endorses corporealism, since she conceives of basic actions 'as actions that are operationally basic' (p. 167), that is, as those actions that one knows how to do 'just like that' or 'by doing no more than electing to do it' (p. 168). She cites, as examples, actions like moving our arms, picking up the cup just before us, writing familiar words, or playing familiar tunes (p. 167). I think that the FPA thesis is implausible because, while we are first-person authoritative over these actions under the why-description, we do not seem to have any special, asymmetric epistemic authority over them under the what-description. This is suggested by a variety of cases of motor illusions. For example, the motor illusion apparently common to anaesthetic patients. William James (1981 [1890]) observes that if you close the eyes of one such patient, hold his anaesthetic arm still, and instruct him to raise his arm to his head, he will be astonished, upon opening them, to find that the movement has not occurred. James says that all reports of anaesthetic cases seem to mention this illusion (p. 749).

A question that arises here is: how can a corporealist version of the FPA thesis explain the shock in the eyes of the anaesthetic patient (upon realising his failure of self-knowledge)? Perhaps one of the immediate responses from O'Brien will be to dismiss this case as irrelevant to the FPA thesis, which concerns only the normal or central cases - that is, cases in which no delusion, no self-deception, and no error are involved. In response, I do not see any valid justification for the consideration of only the so-called normal cases. A standard scientific methodology for understanding the workings of any system of the human body or mechanism of the human mind

is through an examination of the pathologies to which they are susceptible (Levy and Bayne, 2009, p. 459). For example, psychologists have studied visual illusions to understand the operation of the visual system, and motor illusions for insights into the operation of the proprioceptive system (Jones, 1988, p. 72). And so we can equally examine our failures of self-knowledge, or the pathologies to which our self-knowledge capacity is susceptible, to delineate the nature and source of our practical knowledge.

Here, such an examination involves making sense of the anaesthetic patient's astonishment. Two lessons can be drawn from it, both of which invalidate a corporealist construal of the FPA thesis. First, the fact that he was shocked to realise he mistakenly believed that he had lifted up his arm suggests that we do not have first-person epistemic authority over the (supposedly) basic actions embodied in our physical movements. The anaesthetic patient case suggests that my claim to first-person epistemic authority over my arm's movement is not on secure grounds. Second, an intuitive explanation of the patient's shock is that he felt or believed that he had *done* something - what he usually does - to execute his intention to raise his arm. If he had not done anything, he would not have had that shock in his eyes. What could he possibly have done then? Could it be that he had 'elected' or intended to raise his arm? A decision or intention is an unlikely candidate for explaining the anaesthetic patient's shock. We are not normally shocked if our limbs do not move upon our deciding or intending to move them. I think that a natural, more plausible answer to the question raised above is that the patient must have tried or willed to lift up his arm. If so, then the act of trying or willing to do something is what deserves to be described as a basic action. Something that we can do 'just like that' does not seem to be something that we can fail to do. We can fail to move our bodies, but it does not seem that we can fail to try to move our bodies. But if basic actions really consist in such mental events as tryings or willings - that is, if volitionism is true -, then the FPA thesis is trivially true.

1.2. Now is practical knowledge independent of perception or investigation (under both the why-description and the what-description)? The idea of independence means here that our practical knowledge is neither acquired nor justified via perceptual or observational evidence. This information, both about ourselves and about the world, is only a precondition of practical knowledge. It forms part of the background that makes our practical knowledge possible, but we do not need it in order to know what we are doing when we act intentionally (Velleman, 1999;

O'Brien, 2007; Campbell, 2015). Consider, for instance, the case of your opening a window and your knowledge of it. For you to know non-observationally that you are opening a window, Velleman claims, you need to have past observational knowledge of something that is a window and of your ability to open it. But at the (present) time of performing the action of opening the window, no further observation is required for you to know that you are opening it (Velleman, p. 20).

The thesis that practical knowledge requires perceptual or observational knowledge only as a precondition seems to imply that an individual who has lost all his perceptual abilities after years of successfully using them should nevertheless be able to know non-perceptually what he is doing. This sounds very implausible. Consider a complex behavioural sequence such as playing a piece of piano music. Suppose you are a perceptually (both aurally and visually) impaired virtuoso pianist and you are seated before a piano and asked to perform an all too familiar piece. Or suppose that you are a perceptually (visually) impaired person guided by a dog and that part of your all too familiar journey to work at the University of Sheffield involves walking up High Street. If the thesis under consideration has any plausibility, you will be able to know non-observationally what sequence of the piano piece you are playing, or what part of the journey (say, High Street) you are walking, at any time. I think that common sense will have it here that this sort of knowledge is very unlikely. Not only common sense, but even science, I may add.

But O'Brien claims that evidence from experimental psychology confirms that practical knowledge is independent of perception and investigation. She adduces four pieces of such evidence. First, 'subjects are typically aware of acting before their bodies move' (p. 168). Consider my action of raising my arm (for one reason or another). This action consists in the movement of my arm going up under my control or at my command. If I am typically aware of raising my arm before my arm moves, what am I exactly aware of? Certainly not of the movement of my arm, which has not yet occurred. If anything - I mean anything action-like -, I can possibly be aware only of my trying or willing to raise my arm, where this trying is an internal or mental event rather than an external or physical one. This is probably why James's anaesthetic patient was shocked when he opened his eyes and saw that his arm had not gone up. If so, then what we are typically aware of before our bodies move is our trying to move them.

But this awareness does not validate the claim that our knowledge of our bodily movements is non-perceptual. It is awareness of a mental event, and not of a physical event.

Second, ‘subjects who lack bodily awareness can be aware of acting with observation’ (p. 168). As it stands, this ‘evidence’ clearly disfavours the thesis that practical knowledge preempts perception and observation. Did O’Brien perhaps mean to say that these subjects ‘can be aware of acting *without* observation’? If so, then an explanation (of this non-observational awareness of action) that does not appeal to perceptual or sensory inputs is in order here. But O’Brien does not provide any explanation at all of how the ‘evidence’ she has picked supports her thesis that practical knowledge is non-observational.

Third, ‘subjects who are paralysed or lack body parts can take themselves to be acting, in the normal way, even [though] they are obviously not receiving perceptual feedback from a moving limb’ (pp. 168-9). Put simply, (it is claimed that) these subjects can form a non-perceptual belief that they are moving their paralysed or missing body parts in the normal way. A clinically puzzling case that comes immediately to mind here is the phantom limb syndrome, as a case of missing body parts, and its corresponding illusions. But how is it that these patients ‘can take themselves to be acting’ normally? And what does these people’s ‘acting in the normal way’ exactly amount to? One of the most comprehensive studies of the phantom-limb illusion, using 300 amputees in prison-of-war camps during World War II, found out that movements of the phantom-limb ‘are always consciously made and do not occur spontaneously’ (Jones, 1988, pp. 76-77). This fact is taken to suggest that these patients’ sensations of limb movement or, so to speak, their taking themselves to be acting in the normal way, ‘can result from the voluntarily generated motor command’ (ibid., p. 77).

In volitionist terms, we may say that their sensations can result from internal occurrences of tryings or willings. As James also observes,

many of these patients can *will the foot to move*, and when they have done so, distinctly *feel the movement to occur* [...]. Of all possible experiences, it is hard to imagine any pair more uniformly and incessantly coupled than the volition to move, on the one hand, and the feeling of the changed position of the parts, on the other [...]. The process of the willing ought, then, to pour into the process of feeling the command effected, and ought

to awaken that feeling in a maximal degree provided no other positively contradictory sensation come in at the same time (p. 750, emphasis in the original text).

This (speculative) psycho-physiological explanation (of how phantom limb patients come to take themselves to be acting normally) concurs with the interpretation of empirical evidence presented above. And both the explanation and the interpretation appeal to willings or centrally generated motor commands and the resulting sensory or perceptual feedback to explain the phantom limb patients' (illusory) sense of acting normally. From the fact that these patients 'are obviously not receiving perceptual feedback from a moving limb,' as O'Brien notes, it doesn't follow that their taking themselves to be acting normally is not grounded in any such feedback or sensation whatsoever. Their sense of acting normally is perceptual, and this perception 'has been shown to depend on the presence of innervated muscle[s] in the stump' (Jones, p. 78).

Finally, the fourth piece of empirical evidence adduced by O'Brien is that

subjects suffering from the phenomenon of Anarchic Hand, and who are not aware of seemingly purposive movements of their bodies as their actions, typically have reliable perceptual (including proprioceptive) feedback with respect to the movements of their bodies, but *are* damaged in that area of the brain taken to be responsible for motor specifications (p. 169).

The first thing to note here is that it is not immediately clear how the experience of patients with alien hand syndrome (AHS) can be used to buttress a claim about non-perceptual knowledge of actions. O'Brien does not discuss any of the pieces of evidence she presents; she simply lists them. One way that we may understand O'Brien's fourth empirical evidence is as stating or meaning that AHS patients are not aware of their bodily movements or their actions in spite of receiving reliable perceptual information about them, and as suggesting, accordingly, that practical knowledge is not epistemically grounded in perception. But this would be a very strange thing to say about AHS. This pathology is generally used to understand the functioning and disruption of volition and the sense of agency. It is 'a neurological condition in which patients experience their limb movements as unaccompanied by any sense of agency' or as involuntary (Fried et.al., 2007, p. 10844). These movements, of which the patients may or may not be aware, are thus experienced as mere happenings and not as actions, because they are not

initiated - and cannot be guided or stopped - by the patients. So it is absolutely impossible that AHS patient experience will support claims about non-perceptual knowledge of actions. Because there are no actions here in the first place.

1.3. The third and last claim of *exceptionalism about practical knowledge* is that practical knowledge is immediate/spontaneous. These two properties are sometimes used synonymously in the literature, but there is a clear sense in which they are rather semantically different. Proponents of exceptionalism like Campbell characterise practical knowledge as being psychologically immediate, by which they mean that we are typically able to say ‘straight away’ what we are doing intentionally - without having to think about it, notice it, discover it or figure it out (Campbell, p. 8). But Velleman, another proponent, uses the predicate ‘spontaneous,’ by which he means that practical knowledge seems to be knowledge that we originate or invent (1999, pp. 24-5). This feature is particularly tricky to characterise precisely. We may resort to a contrast with the property of immediacy as a first step toward illuminating it. Consider the sort of knowledge you typically have of your stomach pains: when you suffer these pains, you usually know it, without inferring it from anything else. That is, you typically have an immediate knowledge of these pains. But immediate knowledge is not spontaneous knowledge: immediate knowledge of pains is, in a sense, caused by them. By contrast, practical knowledge as spontaneous knowledge is, in Aquinas’s idiom, ‘the cause of what it understands’ (Wiland, 2012, pp. 123-4). Or, as Velleman (1999) puts it, your knowledge of what you are doing is spontaneous because ‘the action was your idea or what you had in mind’ (p. 24).

Let us put aside the feature of spontaneity (because of its lingering mysteriousness), and attend only to the property of psychological immediacy. This property strikes me as the single unproblematic feature of our practical knowledge (so long as it is decoupled from the pre-emption of perception and inquiry). We generally know immediately what we are doing intentionally, both under the why-description and under the what-description. Except, of course, when we have a lapse of memory, as it often happens. In those cases, what normally escapes us is the why-description of our actions. And also people with certain (rare) conditions may not always know immediately what they are doing. Consider the case of Gary. His mind has all its faculties functioning, but he is afflicted with an extreme and intractable case of brain aneurysm sustained a few years earlier. His condition is such that his memory systematically flickers on

and off exactly every five minutes or so. When his memory flickers on, Gary can acquire and hold new memories. During this time, Gary's other mental capacities and systems can operate normally too. And when his memory flickers off, everything stored in it is wiped off, and so Gary's memory turns into a blank slate. Obviously, Gary's enjoyment of epistemic immediacy will be severely limited. Cases like this and others pose problems for the claim that our practical knowledge is not acquired or justified in any of the ordinary ways.

Now, much as the thesis that our practical knowledge is psychologically immediate is very plausible, or even true, it does not pick out anything exceptional about our practical knowledge in any obvious or meaningful way. In addition to attaching to sensory states (like pains), as I mentioned earlier, psychological immediacy also applies to psychological phenomena like visual, auditive or proprioceptive perceptions. As Proust (2005, p. 167) rightly points out, typical cases of visual or auditive perceptions also have an immediate access to their objects. We normally do not (need to) infer from any sensations that we are seeing a white cup on the dinner table or that we are hearing the hissing sound of the pressure cooker before us. We perceive these things immediately. So this thesis being true does not support exceptionalism about practical knowledge.

1.4. To conclude this discussion of exceptionalism about practical knowledge, it appears that this thesis is implausible. We do not seem to know our actions in a special way: our practical knowledge is not first-person authoritative, and it is not independent of observation. If anything, these features apply only to the why-description, and not to the what-description, under which we normally know what we are doing intentionally. Now it is true that our practical knowledge is psychologically immediate, but psychological immediacy also attaches to mental phenomena such as sensory and perceptual states. So this property is surely not distinctive of our practical knowledge.

I would like here to briefly try and explain away the apparent features of first-person authority and preemption of observation or enquiry. It seems that the view that practical knowledge is first-person authoritative and independent of observation springs from a certain kind of psychologism about actions. By the term 'psychologism' I mean 'the mistake of identifying non-psychological with psychological entities. For instance, philosophers who think that logical laws

are not psychological laws would view it as psychologism to identify the two' (Kusch, 2024). Instances of psychologism about action are, for example, O'Brien's identification of actions as psychological phenomena (p. 169), and Velleman's explanation of the supposed spontaneous character of our knowledge of what we are doing by the claim that 'the action was your idea or what you had in mind' (1999, p. 24). Psychologism about actions is certainly false because most of our actions are events in the world, and only a few are mental occurrences.

2

I have just argued that we normally know what we are doing under both a why-description and a what-description, and that this knowledge is not first-person authoritative and independent of perception, but only (though not distinctively) psychologically immediate. That is, the knowledge that we typically have of our actions seems to be acquired or justified in the ordinary ways. But what are these ways?

Actions are purposive events or movements under the control of their agents. Acting is, or essentially involves, exerting control over our selves (body and mind) and over the world for some end. Accordingly, there are three collaborative or inseparable sources of knowledge of the entire range of our intentional actions: the body, the mind, and the world. It is the interaction of these sources and the integration of the data that they yield that provide us with the knowledge of our actions.

Information from, and about, the body is required because most of our actions are obviously physical movements. One and the same physical movement can embody a variety of intentional actions, depending on the context of performance and the underlying intention. For example, the same movement of my leg can be my action of kicking you or my action of just stretching my leg. Whether my leg movement is a kick or a stretch is, at least partly, a function of my intention for making it, which is something in my mind. From the mind come components of our actions that we typically know non-inferentially and non-observationally. These items - notably, our motivational states and other action-explaining states - are elements of the why-descriptions under which we know what we are doing. But our actions are not defined and shaped only by our minds and bodies. They occur in the world and are constrained and structured by it. For

instance, part of what determines whether a certain movement of my arm constitutes the action of ‘punching you’ or ‘waving to you’ is your (real or imagined) presence in front of me in a shared space or world. That is why external evidence is a component of our knowledge of our physical actions. We need visual information about our bodies in motion and other information from the world to know what we are doing.

In this section, I want, however, to assume the truth of *exceptionalism about practical knowledge*. I want to assume that the knowledge that we have of our intentional actions is special in order to examine Velleman’s account of the possibility of such knowledge. Recall that Velleman hopes that if we understand our ‘special’ way of knowing what we are doing we could understand what it means to do something. (In fact, Velleman (1999, 2000) and other philosophers who belong to the Anscombean tradition in action theory (e.g. Setiya, 2007) claim that one of the problems of the standard (or desire-belief) theory of action is that it cannot account for our special practical knowledge. And that, because of this failure, it cannot capture human agency.) Now Velleman argues that practical knowledge is constituted by, or embodied in, the agent’s intention, where this is a kind of belief or a cognitive attitude. Following Campbell (2016), I’ll call this view *cognitive intentionalism about practical knowledge*.

But to understand how intention embodies practical knowledge on Velleman’s account, we need to grasp Velleman’s conception of the nature and structure of the will. Velleman conceives of the will as a desire-driven faculty of intentions. More precisely, it is a composite of intellectual and motivational components: the capacity to form intentions (which embody our knowledge of what we are doing), and the desire to know what we are doing or the drive for self-knowledge. Now how is practical knowledge possible? Velleman holds that, as self-conscious and inquisitive creatures, we are driven to know what we are going to do and why. This drive, which he claims is functionally identical with ourselves as agents, refrains us from performing an action until we are prepared to understand it, that is, until we have formed an idea of it and its supporting reasons or motives (2007a, 2007b). Forming and enacting such an idea is, according to Velleman, what ‘the process of forming and acting on an intention’ amounts to. The intention will then embody practical knowledge, because we have a broad disposition towards self-consistency - that is, a tendency to behave in ways that cohere with our intentions -, a tendency incentivised by our drive for self-intelligibility (2007a, pp. xvii-xix). Velleman thus explains the

possibility of self-knowledge by means of the interplay between the drive for self-understanding, self-consistency as a disposition rooted in this drive, and intention as a cognitive attitude.

In what follows, I call into doubt the postulation of the drive for self-knowledge as a necessary ingredient of practical knowledge (2.1), and adduce some reasons why Velleman's conception of intention is implausible (2.4). In between these sub-sections, I examine what Velleman has presented as the empirical-psychological foundations of his variant of cognitive intentionalism. He claims that social psychologists cite people's tendency to fulfil their intentions, or the antecedently framed conceptions of their actions, to explain how people know what they are doing, and that they give this tendency the label 'act identification' (2007a, p. xviii). I examine the literature on action identification theory, and argue that this theory does not support Velleman's view of intention, intentional action and self-knowledge (2.2). Velleman has also argued that psychologist Prescott Lecky's self-consistency theory holds self-consistency to be driven by the drive for self-understanding. Using textual evidence, I argue that Velleman misinterprets Lecky (2.3).

2.1. Let us begin with the drive for self-intelligibility. If at all we have this drive, does it play any role in our acquisition of self-knowledge? It is a hard fact about us that we are wondering creatures. Wondering reveals a drive to know or understand. As Aristotle states at the beginning of his *Metaphysics*, every human being has a natural desire to know, as evidenced by 'the delight we take in our senses,' and more especially the sense of sight (I.I, 980321-7). This epistemic desire is presumably one of the fundamental human desires. If we did not have this drive to know, there certainly would be no science and philosophy, no research and education, no teaching and learning, and no dating and socialising in the world. We can direct this drive at just any object, including our own selves, which we sometimes do. But does this imply that we have a fundamental drive for self-understanding, that is, 'a fund of labile psychic energy' that motivates our various efforts to make ourselves intelligible (2007a, p. xx), as Velleman claims? And is this alleged drive crucial for the sort of knowledge that we have of our actions?

To the first question I believe we are justified in replying in the negative. One reason for this is that we do not seem to be characteristically and constantly at a loss about what we are doing. Velleman cites episodes of a lapse of memory and the resulting unease as revelatory of our

(abortive) drive for self-knowledge. But given how rare or infrequent such episodes are in the lives of most of us, which Velleman recognises, why should we believe that the drive supposedly manifested by them is fundamental? It would be very odd to posit a specific fundamental intellectual drive behind every specific (kind of) item that we may happen to forget on a given occasion. For example, if you fail to remember the name of the street on which your family lived when you were starting primary school, it would strike you as ridiculous if someone told you that this failure revealed your fundamental drive to know the name of that street. You would be inclined to believe instead that your memory failure revealed just an occurrent, rather than a standing, desire to know that item.

Another reason for thinking that if we have a drive for self-understanding it cannot be a controlling drive is that it is not clear or obvious what would be badly lacking in our lives if we did not happen to have it. Velleman claims that our drive for self-understanding operates like our abiding desire to avoid pain: it continually guides our conduct but only subliminally, without our being consciously aware of it. The desire to not get hurt influences our everyday behaviours like brushing our teeth, walking down the stairs, chewing food, etc. Yet this desire rarely intrudes into our consciousness as we perform these actions. It influences us in silence (1999, p. 37). True. The desire to avoid pain is a vital and controlling desire. But can we say the same about the drive to know what we are doing? One thing is to hold that both drives are subliminal, which is not in question here. Another is to claim that they both have a controlling role in our lives, which is what is in doubt here. My contention is that the drive for self-understanding is not comparable to our standing desire to avoid pain in this regard.

If we did not have a desire to avoid pain, which could result from a congenital or acquired inability to feel pain when we are exposed to harmful external stimuli or when our internal organs are damaged, the integrity of our bodies would surely be seriously imperilled. Congenital insensitivity to pain is found in patients suffering from the extremely rare condition known as congenital analgesia, in which a person never feels pain. And acquired insensitivity is characteristic of, for instance, leprosy patients. These patients are said to experience no pain in the cooler peripheral bodily parts (such as the digits and the nose) because of the destruction of noxious stimulus detectors in those parts by leprosy bacilli. Now both types of patients are generally reported to exhibit pathological changes in their bodies, as a result of their being

deprived of the natural self-protecting systems (Grahek, 2007, pp. 7-9). So the capacity and desire to avoid pain is of a central biological importance in our lives. Can anything analogous be said in favour of the drive for self-understanding? Perhaps this drive is epistemically crucial for practical knowledge as the desire to avoid pain is biologically crucial for bodily integrity?

Let us now examine the issue of whether the drive for self-understanding is crucial for practical knowledge. Suppose that you are attending an auction at Auction House London (- the example with which I started my discussion of exceptionalism about practical knowledge). Suppose, further, that at some point you decide to enter a bid for an antique alarm clock, and you believe that you can do so by just raising your arm. And so you raise it. I will assume that this is one of those actions of which an agent can have the sort of special knowledge that Velleman describes: knowledge that is spontaneous and independent of observation and enquiry. Now how can you know spontaneously and non-observationally that you are raising your arm during the auction? I would like to introduce a little complication here: you have an extraordinarily rare congenital cognitive disorder: your drive for self-understanding mechanism is congenitally inoperative in your mind, quite like the pain systems - that is, the avoidance system and the restorative system - are dead in a patient with congenital analgesia. But, as if to compensate you for this injustice, Nature has endowed you with an exceptionally powerful working memory: it never fails, so you never have any lapses - say, temporarily forget the goals or purposes you are pursuing at any time. Would you, in spite of this congenital defect, still know both spontaneously and non-observationally that you are raising your arm?

Velleman's answer must, of course, be negative insofar as he holds that a standing desire for self-understanding is a necessary ingredient of our special practical knowledge. But why would it be impossible for you to have such knowledge if you had a dead drive for self-understanding mechanism? Recall that you typically know what you are doing both under a what-description (e.g. raising your arm during an auction) and under a why-description (e.g. raising your arm in order to enter a bid for an antique alarm clock). So at least one way to prove your special practical knowledge here would be to answer spontaneously and non-observationally the questions 'what are you doing?' and 'why are you doing it?' should anyone ask them. It seems to me that you would answer them perfectly correctly in spite of your strange cognitive condition. Your perfectly functional proprioceptive system - since your arm has never failed to go up when

you try to raise it, and you have no reason to suspect that it might fail on this occasion - would enable you to know spontaneously and non-observationally that you are raising your arm. And the active presence of your relevant occurrent motives in your sublime working memory would enable you to know spontaneously and non-observationally that you want to enter a bid for an antique alarm clock.

So if, upon seeing your arm go up, your friend sitting next to you asks 'what are you doing, Paul?', you would simply reply that 'I'm entering a bid for that antique alarm clock'. And you would do so spontaneously and non-observationally, in spite of having a dead drive for self-understanding mechanism. Now Velleman must dismiss this explanation. And one way to do this would be to contend perhaps that our motivational and memory mechanisms could not be functional without the subliminal operation of a drive for self-understanding behind them. But should we believe this hypothesis? Invoking some mysterious fundamental drive to explain our practical knowledge amounts to postulating a 'horse in the machine' like the villagers in Paulsen's story that R.E. Hobart reports in his famous paper on free will. In this story about the introduction of the railway into Germany and about German villagers' first sight of a locomotive engine, the villagers are given the clearest explanation possible of how the engine works, but they remain convinced that there is a horse inside doing all the work (Hobart, 1934, p. 3). But there actually is no horse inside a locomotive engine. And there seems to be no horse in the engine of our capacity for practical knowledge.

Velleman might still insist that there is truly a horse in the machine of our self-knowledge capacity, and provide, as evidence, our broad disposition towards self-consistency, that is, our tendency to behave in accordance with our intentions or self-conceptions. Velleman says that psychologists have provided abundant evidence for this tendency and cite it to explain our self-knowledge: 'people know what they're doing because they tend to do what they have just now thought that they are just about to do' (2007a, p. xviii). The psychologists that allegedly relate self-knowledge to self-consistency and intention and that Velleman refers to here are the authors of action identification theory. Now Velleman also argues that psychologist Lesky's self-consistency theory holds self-consistency to be driven by the drive for self-understanding. So Velleman might stress the presence of a horse in the engine of our self-knowledge on the ground that, according to action identification theory, self-knowledge is possible because of our broad

disposition towards self-consistency (2.2), which, in turn, is, according to Lecky, driven by the (primitive) drive for self-understanding (2.3). Let us examine these claims.

2.2. Let us start with action identification theory and Velleman's claims about it. Velleman makes these three claims about this theory: (i) the theory cites our tendency to behave in accordance with our intentions - or our self-consistency - to explain our self-knowledge; (ii) its authors label this tendency or mechanism 'act identification'; and (iii) 'they invoke this mechanism, not only to explain how people generally know what they are doing, but also as a model for the process of acting on an intention' (2007a, p. xviii). Does action identification theory hold these views? And does it support Velleman's cognitive intentionalism about practical knowledge?

To address these questions, we need first to know what action identification theory is about. Developed by social psychologists Vallacher and Wegner (1985, 1986, 1987, 2011), action identification theory is a cognitivist account of the mental control of action, or an account of the causal relations between people's cognitions and their actions. It starts off from the assumption, or fact, that people know what they are doing: they have in mind ideas, descriptions or identities of the actions they are performing. The theory then proposes that a person's action may have various identities and that these identities tend to be organised in a hierarchical manner 'in the person's cognitive representation of action' (1986, pp. 19-20). In the hierarchy, lower-level identities 'convey the details or specifics of the action and so indicate how the action is performed' and higher-level identities 'convey a more general understanding of the action, indicating why the action is done or what its effects and implications are' (1987, p. 4). For example, 'moving one's arms and legs' is lower in level than 'stepping on pedals and turning the steering wheel' because the former indicates how the latter is done. Similarly, 'stepping on pedals and turning the steering wheel' is lower than 'driving a car,' which, in turn, is lower than 'going to work'. 'Driving a car' is higher than 'stepping on pedals and turning the steering wheel' because it indicates why 'stepping and turning' is done, and 'going to work' is higher than 'driving a car' because it suggests why 'driving' is performed (2011, p. 33; 1986, p. 20). Now we may say that all this talk of our knowledge of our actions under hierarchical identities boils down to the simple claim - repeated throughout this chapter - that we normally know or identify our actions under a what-description and a why-description.

From the premise that people know what they are doing under descriptions or identities that are hierarchically arranged in their mind action identification theory goes on to formulate a set of three principles of mental control of action or self-regulation. In fact, this theory is characterised, by its authors, as ‘a system of three principles explaining how people’s thoughts of what they are doing relate to what they do’ (2011, p. 327). These principles are: first, people maintain an action in accordance with their prepotent identity for it. People’s prepotent identity for an action is their ‘idea of what they are doing or want to do (1987, p. 4) and it ‘functions as an intention to initiate [the] action, a frame of reference for performing [it], and a criterion to assess how well [it] has been performed’ (2011, p. 331). This principle captures a widely shared conception of the motivational and action-guiding roles of intentions. Second, when an action admits of both a lower-level identity and a higher-level identity, the latter tends to become prepotent. ‘The idea here is simply that people are always sensitive to the meanings, consequences, and implications of what they do’ or that people’s behaviour is generally motivated by their ‘goals, plans, values, and other global constructs’ (2011, p. 332). Third, ‘[w]hen an action cannot be performed in terms of its prepotent identity, there is a tendency for a lower-level identity to become prepotent’ (p. 332). The basic idea is that people tend to change and lower their expectations or goals in the face of external obstacles and personal difficulties (p. 332). For example, a student may initially aim for a first-class degree only to opt for a second-class upper after failing to score first-class grades in a number of modules. Together, these three principles explain human agents’ cognitive representation and control of their actions.

Now does action identification theory hold the views that Velleman attributes to it? This does not seem to be the case. First, there seems to be no textual evidence for Velleman’s claim that this theory cites our broad disposition toward self-consistency to explain our practical knowledge. This theory does not claim that we have a distinctive sort of knowledge of our actions, nor does it make any claim about the possibility of that knowledge. What it claims essentially is that we typically know our actions under hierarchically organised descriptions, and that these cognitive representations shape the initiation, continuation and alteration of our actions. Second, Velleman’s claim that action identification theorists call ‘act identification’ our tendency to act in accordance with our intentions reflects a misunderstanding of the phrase and the theory itself. By the phrase ‘act identification’ the authors mean simply ‘an organized

cognitive representation of the action' (1987, p. 4), and not a behavioural tendency or disposition. Finally, it follows from these two observations that Velleman's third claim about action identification theory - that the theory views act identification as a model for the process of acting intentionally - is false. So it is not true that, according to action identification theory, self-knowledge is possible because of our broad disposition towards self-consistency or acting in accordance with our intentions.

2.3. Let us now turn to Velleman's claim that Lecky conceived of self-consistency as the object of some primitive intellectual drive. One of the main psychological references or 'proofs' that Velleman furnishes for his claim that self-consistency is driven by a motive for self-intelligibility is Lecky's theory of personality as self-consistency presented in his only and posthumously published book, *Self-Consistency: A Theory of Personality* (1945). Velleman claims that 'Lecky regarded self-consistency as the object of a cognitive or epistemic motive from which all other motives are derived' (2006, p. 224). To properly scrutinise this claim, we must break it down into its smaller constituent claims. We may break it down into (i) 'self-consistency is the object of a cognitive or epistemic motive' (Claim 1 or C1) and (ii) 'this cognitive or epistemic motive is the source of all other motives' (C2). Equivalently, C2 can be stated as 'the motive that is the source of all other motives is cognitive or epistemic in nature'.

Now is there any evidence for C1 and C2 in Lecky's theory? I want to dispute these claims as a misinterpretation of Lecky's conception of self-consistency. My concern here will thus be only exegetical; I will not examine Lecky's thesis about self-consistency. For starters, this thesis is well summed up in these statements (also quoted by Velleman):

We propose to apprehend all psychological phenomena as illustrations of the single principle of unity or self-consistency. We conceive of the personality as an organization of values which are felt to be consistent with one another. Behavior expresses the effort to maintain the integrity and unity of the organization (Lecky, 1945, p. 152).

Here, Lecky posits the principle of self-consistency as what drives and explains all psychological phenomena, and he proceeds to explain such phenomena as personality and behaviour in terms of this principle. Personality is a system of mutually consistent values. And action strives to preserve the wholeness of this system. In fact, for Lecky, the need to preserve

the wholeness of this system is the only drive of human action: 'One source of motivation only, the necessity to maintain the unity of the system, must serve as the universal dynamic principle' (p. 152). And insofar as Lecky assumes that '[l]ife and activity are coexistent and inseparable' (p. 151), the principle of self-consistency is the principle of life itself. So Lecky must be understood as making claims about our primary drive in life, much like his predecessors, notably Freud, Adler, and Frankl: Freud posited the drive for pleasure (or the pleasure principle), Adler the drive for power, Frankl the drive for meaning, and Lecky the principle of self-consistency or the drive to maintain a system of consistent values. Another piece of textual evidence for the view that Lecky is making a claim about a primitive driving force of human life and action is this: 'By interpreting all behavior as motivated by the need for unity, we understand particular motives or tendencies simply as expressions of the main motive, pursuing different immediate goals as necessary means to that end' (p. 153). This much is enough to show that Velleman misinterprets Lecky: if self-consistency is the primary drive of human behaviour, it cannot be the object of any other drive or motive whatsoever. So C1 is false.

And C2, which says that 'the motive that is the source of all other motives is cognitive or epistemic in nature,' does not seem to have evidential support in Lecky's text. Velleman cites such sentences as 'The subject must feel that he lives in a stable and intelligible environment' and 'In a world which is incomprehensible, no one can feel secure' (Lecky, 1945, p. 122) as supporting his claim (2006, pp. 224-5). But these sentences do not provide any support to his intellectualist reading of Lecky's self-consistency. To see why, we need to place them in their proper discursive context. Lecky makes these remarks at the conclusion of a discussion of his concept of psychological and physical development as learning to solve problems of adaptation in order to maintain a stable state or unified organisation. It is about psychological development that Lecky makes the observations that Velleman takes to buttress his claim (C2).

Lecky writes that psychological development involves a learning that is fundamentally geared towards eliminating conflicts between the organism's values and attitudes or between its expectations and events in the world. (pp. 118-119). To illustrate the necessity of unified attitudes, Lecky gives the example of Pavlov's dog and his experience of an apparent acute neurosis during an experiment about form discrimination. After having been conditioned to associate the appearance of a circle of light on a screen before him with the imminent presence

of food and the appearance of an oval light with its absence - his saliva would flow in the former case (excitation) but never in the latter (inhibition) -, the dog was presented with an ambiguous figure (combining features of both the circle and the shape). The figure had the effect of inducing both (conflicting) tendencies in the dog (pp. 119-121). This (lack of a unified attitude) resulted in 'disorganization and dynamic chaos' (p. 121) in the behaviour of the animal.

Now as an illustration for the need of harmony between one's expectations and occurrences in the world, Lecky gives the cases of traumatic neuroses, which are 'brought about by sudden fright or shock and attended by symptoms of profound disorganization, persistent anxiety for no apparent cause, and recurrent dreams of the traumatic episode' (p. 122). A curious fact about this condition, he notes, is that 'the severity of the symptoms [...] seems to be determined by the lack of expectation rather than by the shock itself. If the event is foreseen so that the adjustment can be made gradually, the nervous consequences are usually slight' (ibid.). It is against this background that Lecky speaks of the need for 'a stable and intelligible environment' or for a comprehensible world: in such an environment or world, the organism (human or non-human) can 'foresee and predict environmental happenings [...] and prevent the necessity for sudden readjustments,' which 'is an absolute prerequisite for the maintenance of unity' (p. 122) or self-consistency. Nothing here suggests that the drive for unity or self-consistency, as the primary drive in Lecky's view, is cognitive or epistemic in nature. Any such indication would have the rather bizarre consequence of attributing a drive for self-intelligibility even to Pavlov's dog. In fine, Lecky's theory of self-consistency does not support the enterprise of positing a horse in the machine of our capacity for self-knowledge.

Now if self-consistency is not at all the object of some cognitive primitive motive, could it nevertheless be necessary for practical knowledge? Recall Velleman's claim that psychologists explain people's knowledge of what they are doing by appealing to self-consistency (2007a, p. xviii). Intuitively, whether self-consistency is a necessary ingredient of practical knowledge should depend, at least partly, on what it consists in. On the issue of what self-consistency is or involves, Velleman appears to hold two conflicting views. One view treats self-consistency as the 'tendency to behave in ways that cohere with [one's] own conception of [oneself] - of how [one] behave[s] in general and of [one's] motives on particular occasions' (p. xvii). The other view treats it as the tendency 'to do what [one has] just now thought that [one is] just about to

do' (p. xviii). On the first view, if a man who cherishes a conception of himself as a faithful husband and who has always lived as such just now feels a desire for an extramarital affair and acts on it, he behaves self-inconsistently. On the second view, by contrast, this man is self-consistent: he has done what he just now thought or desired to do. But this (rather trivial sense of self-consistency) is not our ordinary concept of self-consistency, nor is it held, I presume, by any psychologist on self-consistency. So I'll address the question above against the first conception of self-consistency only.

Remember the example about you raising your arm during an auction at Auction House London, and your friend asking you 'What are you doing?'. Your friend's question is not an expression of ignorance: he has, of course, seen your arm go up, and he knows what this behaviour means in your current venue - that you are entering a bid for the object at the centre of attention. It is instead an expression of surprise or even bewilderment: he feels that your behaviour is completely out of character. You have always been a miser; you have never showed any interest in antique objects; and you both came to this particular event just out of a mixture of boredom and curiosity. For him, your action is nothing short of self-inconsistent. You know this is what he thinks, and you fully share his assessment of your behaviour. In fact, it just occurred to you, upon the antique alarm clock being exhibited, that you wanted to have it. And so you entered the bid for it. The question here is whether in so acting you could possibly understand what you are doing. If self-consistency is necessary for practical knowledge, then it seems that one could not know what one is doing unless one behaves in a way that coheres with one's conception of oneself or of how one behaves in general. This strikes me as simply implausible.

2.4. At this point, there is only one building block of Velleman's account of practical knowledge left for examination: the notion of the will as a faculty of intentions, where an intention is a kind of belief or a cognitive commitment to the truth of what (say, an act-description) it represents. Is an intention thus construed a necessary ingredient of practical knowledge? Needless to say, for it to be an element of self-knowledge, whether necessary or not, it must be the case that we have such a mental state as this in the first place. Theorising about our agency involves making claims about the sort of creatures that we are and the sort of capacities that we have. A theory of action based on misrepresentations of our nature or our powers cannot be true. So is it true that we have

a capacity to cognitively commit ourselves to the truth of a proposition about some future behaviour?

I want to briefly raise two concerns for Velleman's view of the will or intention. These concerns are relevantly similar to those I raised about Bratman's conception of intention in the previous chapter. The first concern relates to the sort of cognitive attitude that is appropriate for the characteristic commitment involved in an intention. Recall that I argued - against Bratman's position that an agent's intention to A supports, but does not require, the belief that he will A - that an intention as a pro-attitude with a commitment to action must involve a positive cognitive attitude (belief) toward the intended action (see Chap 2, subsection 2.1). Now I think that Velleman, too, fails to identify the right kind of cognitive attitude. Velleman claims that intending to A (where A is an act-description) is, or involves, being in a cognitive state of commitment to the truth of A. Drawing on Anscombe, he holds that, by virtue of its indicative character, a statement of intention such as 'I'm going to take a walk tomorrow evening'

is meant to provide grounds on which the hearer will subsequently know that one is going to take a walk [tomorrow evening], and those grounds depend for their validity on the assertion's expressing one's own knowledge to the same effect. Hence an expression of intention must at the same time be an expression of knowledge - of something known, in other words, by being the content of the intention expressed (2007a, p. xiv).

This interpretation of a statement of intention seems to imply that intentions always, and somewhat miraculously, cause their objects to happen, quite as immutably and causally efficaciously as the laws of nature. Yet not all our intentions translate into actions: the world (unpredictable as it is) and ourselves (non-transparent and unpredictable to ourselves as we often are) frequently interfere with them. As I noted in the previous chapter (subsection 2.1), we often forget our intentions; we often change our mind about our intentions; and (without any change of mind) we often fail to execute them. These facts about our experiences of intentions hold true even for the most self-consistent of us all.

Now it is quite hard to reconcile these facts about our experiences of intentions with Velleman's view of intentions as embodying the knowledge or truth of the act-descriptions that they represent. Presumably, unfulfilled or flouted intentions are all but embodiments of knowledge or

truth. Actually, our ordinary understanding of statements of intention is quite different from Velleman's. Ordinarily, when I tell you that 'I'm going to take a walk tomorrow evening,' you will not take this statement as an expression of my knowledge of, or certainty about, a future action. Typically, you will take it as expressing both my considered desire to take a walk tomorrow evening and my belief that my taking a walk tomorrow evening is probable. Arguably, understanding an intention as involving, at least, a conative state and a cognitive commitment to the likelihood or probability of what it represents chimes in better with our own experiences of forming and sharing intentions.

The second concern about Velleman's conception of the will or intention is the problem of the intentionality and knowledge of spontaneous actions. (I raised an objection from the intentionality of spontaneous actions against Bratman's planning theory in the previous chapter.) Velleman claims that the drive for self-intelligibility that is constitutive of the will or intention refrains us from doing anything until we are prepared to understand it, that is, until we have formed an idea of it and its supporting reasons or motives. Now it sometimes happens that we perform actions spontaneously - that is, without conscious thinking or planning, without having had the time to form intentions or to be aware of our reasons, etc. Think of someone suddenly springing into action and jumping in front of a moving vehicle to save a child. As I noted in the previous chapter, in performing these actions, we seem to be driven by our situational impulses and guided by our perceptions. Moreover, we typically do not experience such actions as being unintentional, let alone as being mere happenings. These actions are problematic for Velleman's theory of intention and intentional action because they are intentional and appropriately known by their agents, and yet they do not have as their immediate causal antecedent a Vellemanian intention.

3

In the previous section, I assumed the truth of *exceptionalism about practical knowledge*, or of the claim that we have a special kind of knowledge of our intentional actions, in order to enquire whether Velleman's account of the possibility of this knowledge is plausible. Velleman develops a form of *cognitive intentionalism about practical knowledge*, which construes our practical knowledge as a product of a complex interplay between the drive for self-understanding, self-

consistency as a disposition rooted in this drive, and intention as a cognitive attitude. My examination of each of these supposed ingredients of human practical knowledge warrants, I believe, the conclusion that Velleman's variant of cognitive intentionalism about practical knowledge is implausible. Now if both exceptionalism about practical knowledge and Velleman's cognitive intentionalism about practical knowledge are implausible, then human agency cannot plausibly be conceived as *a way of knowing* - a view I dubbed *epistemicism about human agency*. But in this final section, I assume the truth of epistemicism and briefly argue that it yields a conception of human agency that seems to exclude genuine self-determination or self-governance in action. This argument centres around Velleman's conception of the will as practical reason.

Velleman's conception of the will purports to be a naturalistic version of the Kantian conception of the will as practical reason. For Kant, the will as practical reason is empty (or free) of natural inclinations (see Chapter 4), and so can tame the passions. If it has any inclination, the object of this inclination is none other than the moral law. As such, the Kantian will is the capacity to act in accordance with one's idea or representation of laws, or prescriptions of the moral law. It is in virtue of possessing such a capacity that we are autonomous agents. For Velleman, by contrast, the will is a faculty of practical reason that is driven by the natural desire, or drive, for self-knowledge. As I pointed out earlier, Velleman's will is a composite of intellectual and motivational components: the capacity to know what we are doing, and the desire or drive to know what we are doing. Following Kant, Velleman believes that 'we are autonomous insofar as our actions are determined by our idea of what determines them' (1999, p. 176). I will argue that the truth of this purely intellectualist or perceptualist view would rule out the possibility of genuine autonomy. Velleman's reconciliation of an event-causal theory of action with the autonomy of the agent thus appears to be, at best, unattractive.

Following Frankfurt's arguments, many philosophers, including Velleman, have come to believe and agree that reflective endorsement is essential for autonomous agency. But they generally disagree over what sort of mental state is really expressive of reflective endorsement. For Velleman, endorsement is not only a manifestation or expression of autonomy but also a vehicle of causal efficacy. Velleman conceives of reflective endorsement as an intention or kind of belief, and wants to show how it can be causally efficacious: the desire for self-knowledge that

is constitutive of the will inclines us to refrain from performing any action until we are aware, or have a representation, of our motives, and then it inclines us to fulfil this reflective conception. If this conception is accurate, that is, if it identifies the action that our strongest motives favour, it redoubles the causal or motivational strength of these motives (pp. 176-7). So intention or reflective endorsement determines which action we perform by adding to the motivational strength of the motives supporting that action. As Velleman puts it, 'my motives influence me twice - once as they are and once as I *perceive* them to be' (p. 179, emphasis added). That is, our actions are determined by our strongest motives, where this strength is partly constituted (in the motivational sense) by our intentions or (accurate) self-conceptions.

I think that Velleman's epistemicism about human agency is not consistent with genuine human self-determination. One important feature of our autonomy consists in the fact, noted by Velleman (1999) and other philosophers, that we do not see our choices and actions as being straightforwardly determined by the strength of our motives (p. 178). Instead, it always appears to us that our choices are determined by ourselves. In fact, we see ourselves as being capable of choosing and performing an action even when it is weakly favoured by our motives. And we are typically able to choose and act even when our reasons for competing alternatives are equally weighed. Now Velleman's epistemicism does not seem to allow for the sort of human self-determination that makes these practical facts possible.

As I noted above, on Velleman's view, our actions are always determined by our strongest motives. We act autonomously when we are (accurately) aware of those motives (pp. 176-7). On this view, autonomy or reflective endorsement seems to consist in a mere tracking, perception, or awareness of the strongest motives or the balance of our motives. And the will seems to be reduced to a perceptual capacity. But is accurate self-perception or perception of our strongest motives a marker of genuine autonomy? Accurate self-perception does not seem to be the right measure of our autonomy. The reason for this is that our awareness, conception or knowledge of our motives does not seem to be something that we can meaningfully determine. It is something that happens to us - much like (perceptual) beliefs.

Perceptions, like beliefs or judgements, are externally determined. Their contents are a function of something given in experience, and so escape our control. They are fixed by the interplay

between the state of the world (how the world is) and the condition of our perceptual/cognitive apparatus (how we are). There is no freedom of perception or awareness. As I am writing these words on my laptop, - assuming that my perceptual and cognitive apparatuses are intact or functioning normally -, I cannot help but perceive the laptop and perceptually believe that there is a laptop in front of me. I cannot help this unless, of course, I change places or switch positions or move away (say, I go to the kitchen to make myself a cup of tea) - which are acts of will and not of perception. Similarly, if acting autonomously essentially consists in acting on an accurate perception of our strongest motives, as Velleman's epistemicist conception of human agency holds, then we do not seem to determine our actions in any meaningful sense.

But Velleman might object that there is more to our autonomy on his conception of agency than what I have just described. In fact, he has argued that, on his view, our autonomy also involves the possibility of changing the balance of our motives (i) by correcting an initially mistaken perception of it (1999, p. 177); or (ii) by consciously believing that we have a different preference and having a second-order desire to have that preference (pp. 181-3). In response to this, it may be argued that both in a corrected perception and in a mistaken one we are passive with respect to, or not in control of, what we perceive or how we perceive it. So a correction of a mistaken perception of motives, or a change of perception, cannot be a true measure of human self-control either. In both cases, the agent's decisions and actions are 'straightforwardly dictated by the relative strengths of [his] desires or by [his] perception thereof' (p. 178). Additionally, Velleman's reply to my objection faces the familiar problem of the authority of the second-order desire to speak for the agent, where first-order preferences cannot. So my objection that his epistemicism about human agency undermines human autonomy stands.

Chapter Four: Agency Within Morality? Korsgaard on the Practical Standpoint, Free Will and the Moral Law

[A] person's identity, her essential self, seems to be evinced in some special way in her actions, so that when we respond to her actions, we are responding to her [...]. [T]here is a question about how the legitimacy or appropriateness of these reactions can be explained by an account of agency in terms of a causal pathway that leads through the agent's mental states (Korsgaard, 2014, p. 195).

If you want to understand and explain human agency, from what vantage point should you approach it? A very common and influential view is that you should approach it from the human agent's practical standpoint - that is, from the standpoint of the agent in practical reasoning. Amongst the proponents of this view the Kantians hold that the practical standpoint is fundamentally a standpoint of freedom. As Searle paraphrases Kant, 'the process of deliberation [...] can only proceed on the presupposition of freedom, on the presupposition that there is a gap between the causes in the form of your beliefs, desires, and other reasons, and the actual decision that you make' (2001, p. 14). One reason for stressing the perspective of the agent's practical standpoint and its putative characteristic freedom is that we are the sort of creatures that are concerned about justifying their choices and behaviours. But what is the nature of this freedom and how does it bear, if at all, on the process of justification?

In this chapter, I examine Korsgaard's account of human agency as self-constitution, on which we constitute ourselves into agents by choosing our actions in accordance with the moral law or the categorical imperative (2008, p. 1). This account is grounded in the Kantian answer to the above question or his argument from transcendental freedom to the moral law. Korsgaard's version of this argument can be reconstructed and presented as a chain of logical entailments: Freedom (from natural sensible inclinations) is a fundamental feature of the standpoint of practical reason and choice - that is, we must choose under the idea of freedom when we deliberate about what to do. If freedom (from natural sensible inclinations) is a fundamental feature of the standpoint of practical reason and choice - that is, if we must choose under the idea of freedom when we deliberate about what to do -, then the categorical imperative is the fundamental principle of choice for a free will. If the categorical imperative is the fundamental principle of choice for a free will, then a free choice is a moral choice. This is Korsgaard's

argument for why the categorical imperative is the principle of choice and self-determination in action.

I examine Korsgaard's argument by proceeding as follows. I first consider Korsgaard's premise that freedom is a fundamental feature of the practical standpoint and argue that it is false, or at least implausible, if freedom (from natural sensible inclinations) is construed as a psychological precondition, a structural framework, or a formal norm of practical reflection (Section 1).

Then, assuming the truth of this premise for the sake of argument, I show the impossibility of a logical derivation of the moral law from it. In this connection, I show that, insofar as it implies that for an agent to be self-determined it is both necessary and sufficient that the incentive of respect for the moral law be operative when he acts, setting the moral law as the principle of autonomy yields results that are perversely counterintuitive and inconsistent with true self-determination in action. I also examine Korsgaard's alternative argument - namely, practical reasoning is all about finding unity and choosing in accordance with the categorical imperative is what unifies an agent -, and dispute the unity claim (Section 2).

In the last part of the chapter, I return to the problem of the moral law's conflict with genuine self-determination and diagnose that this problem springs mainly from Kant's de-naturalisation of the free will: Kant emptied the will of all natural inclinations, introduced the moral law in this emptiness, and identified the presence or operation of this law with the will's freedom. I show very briefly that *re-naturalising* the free will solves the problem of genuinely self-determined agency. And, interestingly enough, it also seems a more plausible solution for the two problems for which Kant needed a concept of free will: the problem of showing that the moral law is not a mere mental fantasy but a rational necessity, and the problem of the motivational force of the moral law (Section 3).

1

Korsgaard claims that freedom is a fundamental feature of the standpoint of practical reason and choice (1996a, p. 163). In this section, I try to make sense of this claim by spelling out and examining three senses in which freedom could be understood as a central feature of the

practical standpoint. These senses are: first, freedom as a psychological precondition of practical reflection: we must have a free will or be able to choose independently of our motives in order to deliberate (1.1). Second, freedom as a structural framework of practical reflection: we are in a condition of freedom when we deliberate (1.2). And, third, freedom as a formal norm of practical reflection: we must deliberate and choose as if we were free (1.3). I provide arguments against each of these positions, thereby suggesting that there is no principled reason why freedom should determine how the process of justification in deliberation is conducted.

1.1. Let us start with the first sense in which freedom might be a fundamental feature of the practical standpoint or the notion of freedom as a psychological precondition of practical reflection. This notion is captured by one of the meanings of free will in Millgram's claim that 'if one's will is not free, then the question of how to figure out what to do loses its urgency' (2001, p. 2). Needless to say, the truth value of this statement depends on what free will means or how it is understood here. Now we can understand it in two ways. One way of understanding free will here is in relation to the compatibility issue. On this issue, if we assume that it is true that the universe is deterministic and that libertarianism or hard-incompatibilism about free will is also true, then the will is not free. And so, on Millgram's view, there could be no need for practical reflection in such a universe. This strikes me as unacceptable. It seems plausible that even in that universe agents, though not enjoying the freedom of the will, could still need to figure out what to do. So on the understanding of free will in relation to determinism, the urgency of practical reasoning does not seem to be conditional on the possession of a free will.

Another way to understand the relation between free will and practical reasoning in Millgram's statement is from the perspective of the relation between agents' motives and their choices in deliberation. From this perspective, Millgram's statement can be interpreted as meaning that if one cannot choose in independence of his motives, then the question of how to figure out what to do loses its urgency. This is the notion of freedom as a psychological precondition of practical reflection. Now does practical reasoning presuppose this sort of free, independent choice?

Katsafanas (2013) argues so. He presents and defends what he takes to be a Kantian argument for the necessity of assuming this sort of freedom from the practical standpoint. On this

argument, an agent cannot deliberate while envisioning his choice will be determined by his motives. Katsafanas claims that the reason for this impossibility is practical: ‘deliberating about whether to A commits one to viewing oneself as having a say in whether one will A. If one’s action were simply determined [by one’s motives,] then one would not have such a say’ (p. 122). So deliberating commits us to viewing our choices as undetermined. But Katsafanas also states a stronger psychological precondition than this: ‘if there is to be any such thing as genuine deliberation at all, *agents must be capable* of choosing in a way that is not determined by their motives’ (p. 121, emphasis added).

Now I think that Katsafanas’s argument for the necessity of assuming freedom from the practical standpoint faces some serious problems. I will mention just three. First, the argument implies that in order to engage in genuine deliberation an agent must (know or believe himself to) be capable of choosing without being determined. But how can an agent determine whether he is able to do this? Many philosophers, including Kant and Katsafanas, believe that we experience our choices as undetermined by our motives. I must confess that I have never had this sort of experience, and that I am not able to fathom out how one can have it. Moreover, such an experience does not seem to be consistent with the fact that we are often deceived about the true drivers of our decisions and behaviours.

Second, it does not follow from the alleged fact that deliberating commits us to viewing our choices as undetermined that we are actually able to choose independently of our motives. To save the claim from this objection we may give it the following apparently more plausible construal: deliberating commits us to viewing our choices as undetermined means that in order to deliberate it is necessary that we *believe* that our choices will be undetermined. This construal, however, raises the question of the reason for the necessity of such a belief. Now one way to explain this necessity is by drawing from the resources of the version of the moral argument for belief in God that Kant develops in *Religion Within the Boundaries of Mere Reason*. Kant’s argument is that

without a belief in God, we could not also believe that in each person happiness will eventually be proportioned to virtue, and the suggestion is that if we did not have this belief about happiness and virtue, we would be disheartened to the degree that our

motivation to moral action would suffer. In this case, it is the *belief* that God exists, specifically, that would prevent the hinderance to moral motivation (Pereboom, 2006, p. 560).

In a nutshell, Kant argues that a belief in God is necessary for the preservation of moral motivation. This belief is necessary not for a theoretical reason, but for a practical one, namely, the nourishment of moral motivation in human agents. Similarly, on the improved construal of Katsafanas's point, it can be argued that the reason for the necessity of the belief that our choice will be undetermined is not theoretical but practical: the preservation of the motivation to deliberate. On this view, we may not be motivated to deliberate if we do not believe in the indeterminism of our choices. Theoretically speaking, it may be true that no choices are actually undetermined. But practically we must believe ourselves capable of undetermined choices if at all we are (to be moved) to engage in deliberation. Here, it is this 'belief itself, and not specifically the truth of its content, that is needed to secure [the] practical goal' (Pereboom, 2006, p. 560) of engaging in practical reasoning. Whether we are actually able to make undetermined choices is irrelevant. On this construal, Katsafanas's claim (that deliberating commits us to viewing our choices as undetermined) means that it is practically (i.e. motivationally) necessary that we believe that our deliberative choices are (or will be) undetermined.

Now even this more plausible construal of Katsafanas's point is false: we do not seem to need any external motivation or persuading in order to deliberate: typically, all we need is some perplexity, confusion or uncertainty. And these mental conditions are all natural to us. Besides, it is perfectly possible for us to engage in deliberation in full appreciation of the fact that our motives will determine its outcome. Indeed, this is even more desirable than otherwise. For example, it is highly desirable for Louise that her commitment to a healthy lifestyle be alive in her mind as she engages in deliberation about how best to deal with stress and pressure from her work, and that it determine the outcome.

Third, Katsafanas claims that we will not have a say in our deliberation of what to do if our deliberative choice will be determined by our motives. I can see no reason why Louise should not have a say in how to conduct herself if her commitment determines her choice of yoga

against clubbing. Louise may want her commitment to determine her choice of what to do and may engage in deliberation with this desire. If her choice ends up being so determined, the resulting action must be one that bears the seal of her self-determination. I think that this point and the previous two are plausible enough to constitute good reasons for rejecting the notion that freedom is a psychological precondition of practical reflection.

1.2. Let us now turn to the second sense in which freedom might be a fundamental feature of the practical standpoint. This is the sense in which freedom is a structural framework of practical reflection or the condition within which this reflection takes place. The notion of freedom as a structural framework of deliberation seems to be one of the ways in which Korsgaard construes the centrality of freedom for practical reflection. To see this, we need first to clarify Korsgaard's claim about this centrality. Her claim is not that freedom is a fundamental feature of the practical standpoint in virtue of the nature of this standpoint as such or in itself. Korsgaard would readily grant that even lower animals also engage in practical reflection and choice - indeed, there's overwhelming empirical evidence for this -, but deny that freedom from natural inclinations is a fundamental feature of their deliberative standpoint. In fact, she believes that a lower animal's 'perceptions are its beliefs and its desires are its will' (1996b, p. 93). Her claim is instead that freedom is a fundamental feature of the practical standpoint in virtue of what is distinctive about the deliberating human agent. And what is distinctive about the human agent is the capacity for self-consciousness.

Self-consciousness plays a foundational role in Korsgaard's theory of practical reason and rational agency. Korsgaard characterises self-consciousness as an awareness of one's own mental states and activities, and as a distinctive property of human mentality that is critical for rational agency. Unlike lower animals, she claims, we human animals can be aware of our own mental states and activities and attend to them. This capacity to turn our attention onto our internal states and activities is also a capacity to mentally distance ourselves from them (1996b, pp. 92-3). And this experience of distance is an experience of freedom. That is why Korsgaard claims that freedom (from desires and other sensible inclinations) is 'the condition in which we find ourselves when we reflect on what to do' (p. 96). So freedom is a central feature of the practical standpoint because it is the condition in which we, self-conscious creatures, find ourselves when we deliberate.

But what does the statement that freedom is the condition in which we find ourselves when we engage in practical reflection really mean? One way to understand this statement is to read it as holding that we typically experience freedom, or we are actually free, from our natural inclinations when we deliberate. But do we typically experience freedom, or are we actually free, from sensible inclinations when we deliberate? I think that this understanding of the centrality of freedom is simply implausible. It seems to be based on an overly simplistic picture of the psychic life of deliberating agents. This view faces many serious objections. Here are just two sets of them.

First, it is not easy to tell whether we are experiencing freedom from our desires or our natural motives in general. As I remarked earlier, we are often deceived about the true drivers of our choices and actions. Some of the motives that might be imperceptibly operative in our reflection and choice are those that Hume called ‘calm passions’ (e.g. the sense of beauty and ugliness, kindness to children, love of life, etc.) in *A Treatise of Human Nature*. Calm passions ‘produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation’. And, because reason ‘exerts itself without producing any sensible emotion,’ these passions are ‘confounded with reason by all those who judge of things from the first view and appearance’ (II, 3, 3). So calm passions operate differently from those that Korsgaard describes in this alleged deliberative experience of freedom from desires: ‘I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn’t dominate me’ (Korsgaard, 1996b, p. 93). And because they are imperceptible, emotionless or non-impulsive, it is hard to confidently talk of experiencing freedom from them.

Second, apart from Hume’s calm passions, there are other psychological phenomena that seem incompatible with the deliberative experience of psychic distance that Korsgaard and others (such as Katsafanas) associate with freedom from desires. For example, the phenomenon of volitional disunity, of the kind Saint Augustine experienced during his conversion and described in Book VIII of his *Confessions*. Augustine speaks of an ‘agony of indecision’ (Sect. 8), in which ‘two wills’ (a spiritual will and a carnal will) struggled inside his soul - one drawing him to God’s love and the other pulling him to his own lust -, so that he neither willed to serve God

nor refused to do it with his full will (Sect. 10). In despair, he ‘tore [his] hair and hammered [his] forehead with [his] fists’ (Sect. 8). He didn’t experience his self as something over and above his conflicting passions; what he experienced instead was that his ‘inner self was a house divided against itself’ (VIII, 8). This is clearly the antithesis of a deliberative experience of (actual) freedom from natural sensible inclinations.

The opposite of volitional disunity is another such phenomenon. Korsgaard seems to describe the experience of psychic division in terms of reason as the distancing subject and desires as the distanced objects, or in terms of desires appearing before the court of reason (1996b, p. 94). But there often exists in people - at least in some people - an effortless harmony or matching between thought and inclinations. When this state obtains in an individual, his deliberative experience might not be one of psychic division and freedom from his inclinations. The same may be true of those of our desires which are not options among others, or mere impulses, but part of the very fabric of our being. They have a powerful - and, in a sense, liberating - grip on us. (Think, for example, of the web of powerful cares and loves that Frankfurt called ‘volitional necessities’.) And so they are not the kinds of desires we can distance ourselves from or experience being free from any time we engage in practical reflection.

Now if it’s not freedom from natural inclinations that we may experience in deliberation, what kind of freedom is it then? I suggest that it is some form of freedom of choice. This may be understood as the power to choose between two or more desirable and possible objects or means to given ends. It is not freedom from global or local determinism, nor is it freedom from the influence of motives or sensible impulses. It is instead just freedom in the mild sense of the positive enjoyment of certain resources - whether psychological, cognitive, physical or material - in virtue of which an agent has reason to believe that it is in his power to engage in practical reflection and to do what he will settle as the course to follow. It is freedom as this enjoyment and this power that, I propose, is presupposed by practical reasoning. It is also this sort of freedom that might be experienced or felt in the course of practical reasoning.

But how deep and significant, if at all, is this feeling? In part of the first chapter of *L’Expérience Intérieure de la Liberté* (1994), Nabert offers what we may call an illuminating phenomenology of the feeling of the freedom of choice in a deliberative setting. He says that this feeling is

frequently present in our mind when we deliberate, when we weigh options and hesitate between this or that option (p. 9). It is like the very atmosphere where our volitional life evolves and that it does not arise outside the sphere of deliberation and decision-making (p. 11). But the feeling of the freedom of choice does not colour all our deliberative experiences the same way or in equal measure. It varies in constancy, intensity and prominence according to a number of relevant factors. I will briefly discuss just three.

One such factor is the nature of the choice to make. The choice may be one that changes the course of the deliberating agent's life or one that keeps him on the same course. When we make choices that reverse the course of our life or make a clean slate of the past or do not spring from settled habits, we may experience the feeling of the freedom of choice constantly and its full intensity. But such choices are rare in our lives. Ordinarily, or most often, our deliberations and decisions concern the execution of long held plans or the progressive realisation of ideals to which we have remained obstinately committed or of the goals and values that we are lovingly and continuously devoted to. In these deliberative situations, the feeling of the freedom of choice disappears or fades in the background, whereas we experience especially the feeling of our capacity or incapacity to will with perseverance what we first willed (pp. 9-10).

A second factor is the perspective from which the agent deliberates. The perspective can be one of self-detachment and insouciance or one of self-attachment and urgency. Nabert proposes that the feeling of the freedom of choice is all the more intense when our imagination plays a greater role in the construction of possibilities between which we will choose. In some respects, we never feel freer than when we are caressing by the imagination the possibilities of action with which we are playing, without having to make an immediate decision about them. On the contrary, when we have to take a decision about vital parts of our character or life, sometimes the feeling of the freedom of choice disappears or ceases to be an essential element of our volitional experience (pp. 10-11). Think of deciding where to go on summer vacation with your family next year as opposed to deciding now whether you should let your comatose child undergo an emergency, high-risk surgery in the next fifteen minutes.

A third, and final, factor is the personality and character traits of the deliberating agent. The more an agent's personality is completely bound by his previous commitments, the more his

character is formed and habits take roots, the less place the feeling of freedom of choice holds in his mind when he deliberates (p. 9). Most of his deliberations and choices will typically concern the execution of his long held projects and resolutions, so that the feeling of the freedom of choice will be almost imperceptible in his mind. By contrast, to an agent whose dispositions are such that he does not feel bound to any plans or decisions every deliberative experience will feel like a new beginning and hence involve an intense feeling of the freedom of choice.

Now does the fact that the feeling of the freedom of choice constitutes the very atmosphere where our volitional life evolves, though with varying degrees of intensity, make it a fundamental feature of the practical standpoint? We may here wonder whether we should give this feeling a privileged status over other affective states - such as powerlessness, despair, panic, etc. - that may cross our minds during our deliberation and decision-making. Does the fact that the feeling of the freedom of choice, unlike other states, is characteristically present in all our volitional experiences make it a fundamental feature of the practical standpoint in the Korsgaardian sense (of freedom from natural inclinations as a structural framework of practical reflection)? It certainly does not. The freedom of choice, as characterised here, is not Korsgaard's freedom from natural inclinations and impulses. Unfreedom in Korsgaard's sense can coexist, and is compatible, with the freedom that Nabert describes. I can be in the grip of a relentless desire and yet *feel* free to choose how to satisfy it. If so, then we have reason to reject the notion of freedom from natural impulses as the structural framework of practical reflection

1.3. Finally, let us consider the third sense in which freedom might be a fundamental feature of the practical standpoint. This is the notion of freedom as a formal norm of practical reflection and hence of rational agency. We find this notion of freedom in Korsgaard's statements such as you must choose and act *under the idea* of freedom (1996b, p. 94, emphasis added) or 'you must choose *as if* you were free' (1996a, p. 162), which are rooted in Kant's fictionalist argument that rational agency requires us to act as if we are free. Here, freedom is a fundamental feature of the practical standpoint not because it is an actual experiential property of deliberation. It is such a feature rather because we must choose *as if* we were experiencing it. 'When you deliberate,' writes Korsgaard, 'it is *as if* there were something over and above all of your desires, something which is *you*, and which chooses which desire to act on' (1996b, p. 100; 2009, p. 72, emphasis added). This does not mean that 'you must *believe* that you are free, but that you must choose *as*

if you were free. It is important to see that this is quite consistent with believing yourself to be fully determined' (1996a, p. 162). And we must choose as if we were free means that the conception of ourselves as having free will must determine what counts for us as a reason for action.

But why must we choose as if we were free in the first place? To show us why, Korsgaard invites you to imagine yourself taking part in a scientific experiment, where you know that everything you will do during the day is 'programmed by an electronic device implanted in your brain' (1996a, p. 162). This device will determine each one of your thoughts but not cause your movements mechanically. Korsgaard supposes that you might be plagued by the thought that every move you will decide must have been programmed, and so end up not making any decision. 'In order to *do* anything,' she concludes, 'you must simply ignore the fact that you are programmed, and decide what to do - just as if you were free' (p. 163). It appears that the reason for choosing *as if* we were free is of the same nature as the reason for *believing* ourselves capable of undetermined choices discussed earlier: it is practical and has to do with motivation.

But is this practical reason good enough? For Korsgaard's thought experiment, and the conclusion drawn from it, to justify the notion of freedom as a formal norm of practical reflection, it must be the case, at least partly, that deliberating agents normally or typically are in the sort of frame of mind that the experiment depicts. I doubt this is ever the case for (most of) us. When Louise, who is experiencing increasing stress and pressure from her work, deliberates about the ways to keep calm and focused that are consistent with her commitment to a healthy lifestyle, there's no reason to suppose that, if she happens to think that whatever she'll decide has been fixed in advance in some way, she'll necessarily be crippled by this thought. And the same is true of the rest of us. So we have no reason whatsoever to deliberate and choose *as if* we were free from our desires, cares or interests. And so there's no good practical reason for freedom (from natural sensible inclinations) to be a formal norm of our practical reflection.

2

In the previous section, I examined Korsgaard's claim that freedom is a fundamental feature of the practical standpoint and argued that it is false, or at least implausible, if freedom (from

natural inclinations) is construed as a psychological precondition, a structural framework, or a formal norm of practical reflection. One of these construals, namely, the notion of freedom as a formal norm of practical reflection, is a premise in the chain of logical entailments whose final conclusion is that a free choice is a moral choice. Recall that the chain goes like this: (1) Freedom (from natural sensible inclinations) is a fundamental feature of the standpoint of practical reason and choice - that is, we must choose under the idea of freedom when we deliberate about what to do (*P*). (2) If freedom (from natural sensible inclinations) is a fundamental feature of the standpoint of practical reason and choice - that is, if we must choose under the idea of freedom when we deliberate about what to do (*P*) -, then the categorical imperative is the fundamental principle of choice for a free will (*Q*). (3) If the categorical imperative is the fundamental principle of choice for a free will (*Q*), then a free choice is a moral choice (*R*).

In this section, I want to assume (for the sake of argument) the truth of *P*, or the validity of the injunction that we must choose as if we were free, and examine whether *P*'s being true logically entails *Q*. Now the *freedom claim* about the practical standpoint - that is, the claim that freedom is a fundamental feature of the practical standpoint - is apparently interchangeable with what we might call a *unity claim* about the practical standpoint, or the claim that deliberation is all about finding unity. 'When you deliberate about what to do and then do it,' writes Korsgaard, 'what you are doing is organising your appetite, reason and spirit into a *unified system* that yields an action that can be attributed to you as a person' (2009, p. 179, emphasis added). From this unity claim about deliberation, like from the freedom claim, Korsgaard goes on to argue that choosing in accordance with the categorical imperative is what unifies an agent. In this section, I examine first the freedom claim (2.1) and then the unity claim (2.2) to determine whether they yield the moral law as the law of choice. This analysis suggests that there is no logical path from freedom or unity to morality.

2.1. Let us begin with the freedom claim. To assume the truth of *P* involves, Korsgaard suggests, assuming the following about us as deliberating agents: we are entirely self-determining or autonomous wills. We have no human inclinations, emotions or desires, but only a pure (that is, sensibly undetermined) reason, which demands ultimate explanations. As such, we cannot choose and act for no reason. We are also in some kind of Original Position. In order to have

reasons for action, we must have a fundamental principle of action (which will determine what considerations count as reasons). Because reason demands ultimate explanation, this principle must be something unconditioned, something that admits of no further explanation. That is, it must be something about which ‘it is impossible, unnecessary, or incoherent to ask why a free person would have chosen it’ (1996a, p. 164). And since we are autonomous wills, we must make this principle for ourselves rather than have it imposed on us from outside. Moreover, as free wills, we are spontaneous: we have no antecedent reason to choose one principle rather than another (2009, pp. 97-8; 1996a, pp. 163-7). Now given these essential attributes of our nature as transcendently free wills, what must we choose as our ultimate principle of action?

My initial contention here is that, as things stand, *anything goes*. As such autonomous and spontaneous wills, we could adopt the principle of, say, pursuing only those objects of the inclinations of the first person with red hair that we will get to know or only those things that are generally disgusting to human beings. This principle does not seem inconsistent with our autonomy, our spontaneity or any other of our properties as free wills. And obviously it is not the categorical imperative or any one of its possible species. If so, the (assumption of the) truth of *P* does not entail *Q*. But Korsgaard rejects the adoption of principles like this by appealing to human nature and its constraints. She might claim that the objects of any random person’s inclinations are not a serious option or that generally disgusting things are not an option at all for human beings. There could not be any incentive for a human being to pursue those things, and yet human beings must have an incentive in order to act (1996a, p. 165). So the principle of pursuing only those objects of the inclinations of the first person with red hair that we will get to know or only those things that are generally disgusting to human beings is ruled out. And so is the ‘anything goes’ reasoning.

I think, however, that Korsgaard cannot logically avail herself of the appeal to humanity. For we, as Kantian free wills, are not human beings; we are impersonal, detached, disembodied entities: according to Korsgaard herself, the subject of the standpoint of pure practical reason is just ‘the [free] will itself [...], and not the whole person’ (1996a, p. 166) with his human sensibility. As such, we have no human nature that can constrain our choice of ultimate principles of action, and this choice need not be sensitive to typical human incentives. So, as things stand, *anything*

goes really: the categorical imperative is not a logical consequence of the assumption that we must choose under the idea of negative freedom.

I will now argue that a similar result obtains even if we grant human nature and its constraints to ourselves as Kantian free wills. Humanising ourselves in this context involves mainly allowing that as such entities we are susceptible to only these two kinds of incentives: the incentive of self-love and the incentive of respect for the moral law. To make things even simpler, let us grant further that this susceptibility entails that our choice of the fundamental principle is reduced to a choice between the law of self-love (the law of personal happiness) and the moral law (the law of duty). Given this, and given the autonomy and spontaneity of the Kantian free wills, claims Korsgaard, the categorical imperative (or the moral law) in its Formula of the Universal Law necessarily follows as the principle of a free will. The reason for this is that, rather than impose an external constraint on the will, the moral law ‘merely says what [the will] has to do in order to be an autonomous will at all,’ which is to choose a law (1996a, p. 166). Conversely, the adoption of the law of self-love constrains the choice of a free will: with this adoption the will ‘departs from its position of spontaneity and puts itself in the service of inclination’ (ibid.). In a nutshell, the moral law preserves, while the law of self-love violates, the spontaneity and autonomy of the will.

I believe that Korsgaard’s defence of the moral law is vitiated by two related inaccuracies. The first inaccuracy lies in the claim that the moral law’s only constraint on the choice of the free will is that it choose only a law - leaving it to the will itself to determine the content of that law -, whereas the law of self-love imposes a substantive constraint. This claim is inaccurate because Kant’s Formula of Universal Law requires choosing more than just something that has the form of a law. It says: ‘Act only according to that maxim *by which you can at the same time will that it should become a universal law*. (G 421, emphasis added). I say that the Formula requires more because laws are not analytically universal. And so universalisability cannot be a mere formal constraint; it constrains the content of the law to be chosen. But the real issue is the question of why the will’s putting itself in the service of desire (by adopting the law of self-love) is a negation of its spontaneity whereas its putting itself in the service of duty (by adopting the moral law) is an affirmation of its spontaneity.

Even more pointedly, we need to ask: why isn't going whichever way the will may go a genuine exercise of its autonomy and spontaneity? Korsgaard will presumably reply that the will's putting itself in the service of desire cannot be a genuine exercise of its spontaneity and autonomy because in so doing the 'will becomes a mere conduit of natural forces' (1996a, p. 168). In other terms, it is debasing for the will to be in the service of desire. I do not agree. Although it is perfectly possible, and even logical, for one to spontaneously and autonomously do something debasing without losing one's spontaneity and autonomy, we need to stress here that not all our desires are self-debasing. And that not every form of self-debasement is vicious and heteronomous. In fact, some forms might even be a nobler exercise of one's autonomy and spontaneity. In Saint Paul's theology, Jesus-Christ's self-debasement or submission to the laws of human nature - 'though he was born in the form of God [...], [he] emptied himself, taking the form of a slave, being born in human likeness' (Philippians 2: 7-8) - appears as the most authentic expression of freedom.

Apart from the imposition of a substantive constraint on the will, the other (and perhaps more important) reason why Korsgaard dismisses the law of self-love is that 'there is no incentive for the spontaneous will to [depart from its position of spontaneity]' (1996a, p. 166). I think this is a logical point; it follows from the very nature of the Kantian free will: this entity, we have said, needs a principle of action, but, insofar as it is spontaneous, it has no antecedent reason to opt for one principle rather than another. For the will to do anything at all (including opting for a principle), it must have a reason to do that, and for it to have a reason, it must have first a fundamental principle of action. It seems to follow that this principle must flow, so to speak, organically or naturally out of the very nature of the will - that is, without the will having to do anything.

Korsgaard dismisses the law of self-love on the (logical) ground that it does not flow this way, so that its adoption by the will is inconsistent with the will's spontaneity. This is perfectly true. What is not true - and this is the second inaccuracy - is her claim that the moral law is the organic law of an autonomous and spontaneous will: 'By making the Formula of Universal Law its principle, the free will retains the position of spontaneity. Or, to put it a better way [...], the free will need do nothing to make the Formula of Universal Law its principle: it is already its principle' (1996a, p. 166). This would be true if all that the moral law requires is that the will

choose a law. But, as I have already argued, it requires a universalisable law, and universalisability does not flow organically out of the nature of the free will. So faced with the property of the spontaneity of the will, the moral law and the law of self-love fall together. We have a *nothing goes* problem here.

The *anything goes* problem and the *nothing goes* problem suggest that there is no entailment relation between *P* and *Q*, between the assumption of freedom in practical reflection and the adoption of the moral law as the ultimate principle of action. If so, when Korsgaard (or Kant) maintains that freedom (free choice) is morality (moral choice), she (he) does not logically deduce morality from freedom but rather *reduces* freedom to morality for theoretical or practical reasons. For example, the fear of the arbitrariness of transcendental freedom (Robberegts, 1964, p. 234) or the need to show that morality (the categorical imperative) is a requirement of reason, and not a mere mental fantasy (Korsgaard 1996a, p. 160).

But this reduction surely amounts to a negation of freedom. If an autonomous and spontaneous will can logically choose (or reject) self-love or the moral law as its principle of action, maintaining that only the moral law is the law of autonomy and spontaneity clearly amounts to denying (free) choice and affirming that all that matters for autonomy and spontaneity in action is respect for the moral law. In other words, it is to affirm that for an agent to be a self-determined agent it is both necessary and sufficient that his actions spring out of his respect for the moral law. This affirmation has some perversely counterintuitive implications.

To see this, let's consider the following five agents with their motivational profiles. Agent A has a motivational structure with only one incentive - the incentive of self-love. This means that the only considerations that can possibly or actually move him to act are those that can effectively preserve or increase his happiness. Another peculiar feature of A's motivational structure is that actions are necessarily triggered out of this structure once the incentive of self-love is operative. And this incentive is mechanistically activated as soon as a prudential computational device in the brain of A has a perception or intuition - which is generally infallible - of what A has (most) reason to do in every situation he finds himself in. Agent B is like A in all respects except that his single incentive is that of the respect for the moral law and, correspondingly, he has a moral computational device in his brain. Agent C is exactly similar to A save that he has the capacity to

refrain from acting even when his appetite is active. Agent D has a dual-inclination structure - with both incentives of self-love and respect for the moral law - and, accordingly, two computational devices in his brain. He is invariably moved to act by the incentive that is activated first, where in many cases this tends to be the incentive of self-love. And Agent E is like D in being endowed with a dual-motivational profile and two computational devices in his brain. But, unlike D, no inclination can move him without his consent; and he can act even when neither incentive is operative. But he has made it his policy that every time his incentives clash, he will always favour the incentive of self-love.

Now which of these agents is free, unfree, or freer than the others? It is intuitively plausible to hold that there is no difference in terms of freedom between Agent A, Agent B and Agent D in spite of the differences in their respective motivational profiles. Agent A and Agent B are as much passive with respect to their single incentive as Agent D is with respect to his dual incentives. The three agents are wantons: Agent A is a Simple Selfish Wanton, Agent B a Simple Moral Wanton, and Agent D a Complex Semi-Selfish Wanton. Comparatively, Agent C is a Free Agent - he is active with respect to his incentive - in spite of being motivationally as mono-inclination as the Simple Selfish Wanton and the Simple Moral Wanton. And Agent E is even freer than Agent C; he truly is an Autonomous and Spontaneous Agent.

But if we stand by the affirmation that for an agent to be a self-determined agent it is both necessary and sufficient that his actions spring out of his respect for the moral law, we will get opposite results. On this view of self-determination, it is surely Agent B that is an Autonomous and Spontaneous Agent, because in him there is no impediment whatsoever to the operation of the moral law. As to Agent C, he will be now a Simple Selfish Wanton (just like Agent A), and Agent E might not even qualify as a Free Agent properly speaking. These are the perversely counterintuitive implications of this view of self-determination.

Remember that I introduced this view as one that is not logically entailed by the assumption (or truth) of the idea that we must choose under the guise of freedom. Interestingly, even if this affirmation was a logical derivation - that is, even if a free choice is logically a moral choice -, it would carry the same damaging implications. To see why, let us consider three more cases.

In his youth, during his college days, Agent F decided to act and live always in accordance with the moral law after reading Kant and being persuaded by Korsgaard that no other way of life made better moral or human sense. Over the years, although he remained committed to this higher calling, he would occasionally call it into question and lapse into the pursuit of self-love, telling himself that it was not awful to do so when he was convinced that it would cause no harm to the world. So like the life of perhaps most people, his life was one of devotion to a noble cause mixed with an internal struggle with his sensible nature (doubt and indulgence). But recently, having developed a peculiar kind of obsessive-compulsive personality disorder (OCPD), he has not been experiencing any such struggle: his OCPD has made him single-mindedly and irreversibly committed to the moral law, so that everything he now does is perfectly consistent with it. Agent G has had almost the same life journey and experience with the moral law as Agent F except that his current morally harmonious and internally peaceful life is the hard-won reward of many years of a moral-law-inspired ascetic discipline in a Benedictine monastery that he joined after college. And Agent H is exactly like the pre-OCPD Agent F: he is still leading a life of devotion and struggle, of ups and downs.

Again, which of these agents is self-determined, not self-determined, or more self-determined? Intuitively, Agent F is not self-determined although all his actions bear the stamp of the moral law. He is not self-determined because there is a sense in which his commitment to the moral law comes from his condition rather than from himself. In his actions he is, so to speak, a mere conduit of the workings of his mental disorder. As such, he is simply a different type of the Simple Moral Wanton. In comparison, Agent H is certainly self-determined, as evidenced by his devotion and struggle, although he is not a wholehearted devotee of the moral law. There is no reason to suppose that he is generally passive with respect to his incentive of self-love or his incentive of respect for the moral law. Finally, Agent G appears to be more self-determined than Agent H. This is surely not because he, unlike H, can no longer err from the moral law, but (partly) because he is wholeheartedly self-involved in everything he does.

But on the view that a free choice is (logically) a moral choice, Agent G and Agent F are equally self-determined, presumably because they both serve the moral law or because this law is at work in both of them in equal measure. And this is a very disturbing result. This damaging implication of the view that a free choice is a moral choice suggests that the operation of the

incentive of respect for the moral law is neither necessary nor sufficient for any genuine sort of self-determination in action. If so, whether an agent is genuinely self-determined should not be a function of which element of his motivational makeup moves him to act, but of the relation in which he stands to that element and the entire structure (more about this in Section 3).

2.2. Now let us turn to the *unity claim* about the practical standpoint. As a recall, this is the claim that deliberation is all about finding unity or ‘organising your appetite, reason and spirit into a unified system’ (Korsgaard 2009, p. 179). From this claim, like from the freedom claim just examined and challenged, Korsgaard goes on to argue that choosing in accordance with the categorical imperative is what unifies an agent. There are three questions that we might want to ask immediately about the unity claim. First, deliberation is about finding unity of what? Second, why does that unity matter, if at all, for agency? And if it matters for agency, how is it achieved in deliberation? Korsgaard’s answer to this last question is, of course, that in order to achieve unity, we must choose and act in accordance with morality (2011, pp. 75-6). But to get to how to achieve unity in deliberation we must first show that it is necessary for agency.

I will argue that unity is not necessary for agency. Here is my argument in four steps. First, talk of deliberation as being about finding unity implies the (possibility of the) disunity of what ought to be unified. It is this (still undefined) entity that ought to find unity during deliberation. So the bearer of unity must logically be some entity that can lose unity or be dis-unified. Second, and in response to the first question above, philosophers generally speak of the unity of the self: that is, they speak of the self as the privileged bearer of unity. Third, there are two different senses of the self: the self as subject of conscious experiences and the self as the structure of a person’s cognitions, motivations and dispositions. As subject of conscious experiences, the self cannot be dis-unified. Having and making sense of conscious experiences depend on the self as subject of those experiences being one and single. But as structure of cognitions, motivations and dispositions, the self can be dis-unified. It is presumably with this sense of self in mind that philosophers of unity generally speak of the unity of the self and the unity of the will synonymously or interchangeably. For example, Frankfurt (2006) explains synchronic coherence or unity of the will or self as ‘a willing acceptance of the motives and dispositions by which one is moved in what one does’ (p. 17). When this acceptance obtains, he says, when ‘we are acquiescent to ourselves [...]’, there is no conflict within the structure of our motivations and

desires [...]. The unity of our self has been restored' (p. 18). And Korsgaard (2006) states: 'I think that the unity of the will or the self depends on a formal version of the categorical imperative' (p. 61). Fourth, neither of these senses of the self warrants the claim that the unity of the self is necessary for agency.

Let me defend the fourth step or claim here. First, the self as subject of conscious experiences cannot be dis-unified, and so it is absurd to speak of deliberation as being about finding the unity of this sort of entity. Some philosophers of unity appear to think that this entity can be dis-unified, and even try to describe how it gets dis-unified. For them, it seems that dis-unification typically occurs in episodes of self-reflection or self-awareness. For example, Frankfurt (2006) speaks of such an episode or of our experience of objectifying the contents of our mental lives in terms of 'our division of ourselves' (p. 4), self-division and self-objectification (p. 5), taking ourselves apart (p. 13), or separation from ourselves (p.18). He holds that this self-conscious experience undermines our wholeness (p. 18). Frankfurt's main point is clearly that reflective self-awareness results in the disunity of the self as subject of conscious experiences (and of the self as cognitive and motivational structure).

And we find a similar talk of disunity of the self (in both senses) resulting from an episode of reflective self-awareness in the following passage by Korsgaard (2014):

[W]hen we turn our attention away from what we perceive onto the fact that what we are doing is perceiving, then there is a way in which our experience loses its unity. What was once simply given to us as a unified environment is now given to us as a kind of heap of perceptions or experiences, and it is now up to us to put them back into a unified conception of the world. In a similar way where once upon a time we always knew what to do in response to a situation, our own possible responses are now given to us as a kind of heap of desires and fears and impulses, and it is up to us to put *ourselves back together into unified agents* [...]. The principles of reason are the principles by means of which we *unify, or reunify, both ourselves and our conception of the world* (p. 207, emphasis added).

In their talk, both Frankfurt and Korsgaard appear to identify our self (as subject) with the objects of our conscious experiences. As these objects, sensory and mental, are plural and disparate, it is easy to get a picture of the conscious self as a divided entity once we identify it

with the inputs of its awareness. But then this picture raises the question of who is the subject of those plural and disparate conscious experiences. The picture, obviously, cannot answer this question, because it comes from a distorted phenomenology of consciousness. Against a similar mistake by Hume I argued in Chapter One that our consciousness has a bimodal character. That is, in the very act of perceiving an object, whether internal or external, I become aware not only of the object but also of the one (myself) perceiving and the act of perceiving. And, as Kant remarked in *Critique of Pure Reason*, this consciousness of ourselves as subject is an awareness of ourselves as the ‘single common subject’ (CPR, A350) of various conscious experiences. By this, Kant probably meant that ‘to think of myself as a plurality of things is to think of *my* being conscious of this plurality, “and that pre-requires an undivided *me*”’ (Brook and Wuerth, 2023).

If so, then the whole talk of the disunity of the self (as subject) during episodes of reflective self-awareness is misguided. And to speak, as Korsgaard does, of our perceptual experiences as losing their unity when we make them the objects of our attention is certainly to misrepresent our psychic dynamics. It is the conscious subject that confers unity on its experiences by processing, and attending to, them. In sum, the self as subject of conscious experiences cannot be dis-unified, and hence cannot be invoked to warrant claims about the necessity of unity for agency.

How about the self as the structure of a person’s cognitions, motivations and dispositions? Unlike the self as subject, the self as structure (or as will) obviously can be dis-unified. Recall Saint Augustine’s experience of volitional disunity mentioned earlier. He spoke of the conflict between his ‘two wills’ (a spiritual will and a carnal will) in terms of his ‘inner self [being] a house divided against itself’ (VIII, 8). This is a clear instance of the disunity of the self as identical with the will. Now it is the unity of the self as a person’s will or as a person’s cognitive and motivational structure that ought to be the target or concern of deliberation. But deliberation can be about finding the unity of this self only if this unity is necessary or matters for agency. And this is our second question. Why does the unity of this self matter, if at all, for agency?

Korsgaard claims that ‘action requires agency, and agency requires unity’ (2009, p. 46) and that ‘the kind of unity that is necessary for action cannot be achieved without a commitment to morality’ (p. xii). She argues that unity is essential to agency, because an action, unlike a mere

happening, is supposed to be a movement effected by, and hence attributable to, the agent himself as a whole, rather than a movement arising from something merely working in or on him. Unity, in other words, is required ‘to ground the special kind of attribution that is distinctive of agency: attribution to the agent himself’ (2014, p. 193) and to ground the reactive attitudes - ‘like love and hate, liking and disliking, gratitude and resentment’ (p. 194) - that often go along with that attribution.

I disagree. I think that Korsgaard’s answer is implausible, and that the unity of the self as a person’s will or cognitive and motivational structure is not necessary for agency or for the attribution of responsibility. Our social practice of attributing agency and responsibility to each other does not presuppose the unity of our cognitions, motivations and dispositions at the time of action. This practice is, of course, sensitive not only to the nature and consequences of actions but also to the mental states of their authors. But it is not sensitive to the unity of people’s wills or cognitive and motivational structures. As a matter of fact, there is a clear sense in which an akratic agent’s cognitive and motivational structure is not unified when he acts: he performs an action that he does not endorse, that he may not identify with or that he may feel alienated from. And yet we normally hold akratic agents accountable for their behaviours. This is an indication that we treat akratic behaviours not as mere happenings but as actions.

The unity of the self as a person’s will or cognitive and motivational structure is not necessary for agency. But we need it. It is the reflexive and reflective structure of our consciousness that imposes on us the need to find unity within the structure of our beliefs, desires and dispositions. As rational and self-conscious creatures, we want our thoughts, desires, feelings, choices and behaviours to hang together. In some obvious ways, this unity makes life, practical reasoning and acting a lot easier. But it is not necessary for agency and responsibility. So the way to the moral law cannot pass through the unity claim about the practical standpoint either.

3

This final section picks up from the conclusion of the discussion of whether the freedom claim (or requirement) entails deliberative choice in accord with the moral law. Recall that my examination of scenarios about agents with different motivational structures led to the

conclusion that whether an agent is genuinely self-determined should not be a function of which element - the moral law or self-love - of his motivational makeup moves him to act, but of the relation in which he stands to that element and the entire structure. I want to show, by means of a medieval theory of free agency, that it is not necessary to denaturalise the will, or empty it of natural inclinations, in order to secure its freedom or self-determination in action, or even to secure the appeal of the moral law. In fact, re-naturalising the will could be the solution.

Now Kant's view that a free choice is a moral choice is a conclusion about human action drawn from his moral theory. But for this presupposition to hold, it must be established independently of the moral theory in which it originated. As Korsgaard (1996a) explained, this view can be traced to Kant's struggle with the problem of demonstrating that the moral law is not a mere mental fantasy but that it presents us with a rational necessity. Freedom enters into the picture of Kant's moral theory as a solution to this problem. In this picture, freedom is meant to connect the fact of being rational and the fact of acting on the moral law as follows: first, a rational person must regard himself as having a free will (or the claim that the freedom is a fundamental feature of the standpoint of practical reason); second, the moral law is the law of a free will; and, third, a rational, free person obeys the moral law, or a rational, free choice is a moral choice (p. 160).

Another, related problem for which Kant's moral theory seemed to need the idea of freedom is the problem of the motivational appeal of the moral law: why would creatures that are subject to their natural inclinations or concerned about their happiness disregard them for the sake of the moral law? Kant's moral theory replies that the motivating thought of the moral law is that freedom or identification with our rational nature gives us a higher vocation - to make a difference in the world - than the satisfaction of our desires, which reduces us to mere conduits of natural forces (pp. 168-9). So Kant resorted to the idea of freedom to secure the rational necessity and the motivational force of the moral law. However, as I have argued, the conclusion that this enterprise reached about free agency not only does not stand but is, paradoxically, a negation of genuine self-determination in action.

Kant denaturalised free will - that is, he emptied it of all natural inclinations - in the hope of securing his theoretical goals. This denaturalisation appears, however, to cause more problems

than it solves. In this last section, I want to show very briefly that re-naturalising the will could be a more plausible option for Kant's moral theory and action theory. By re-naturalising the will I mean making the will natural again, or re-endowing the will with natural inclinations and allowing that the will's freedom is consistent with its having these inclinations. And I am using this expression in reference to the period of time, before Kant, when the (free) will was a natural rather than a noumenal entity. In particular, I am using it in reference to the medieval period and, more especially, to the account of the human will by John Duns Scotus, a 13th century Scottish Franciscan philosopher-theologian.

On Scotus's account, the human will has at least these two essential properties: (1) it is a self-determining power with respect to opposites. And (2) it has two basic inclinations: the (lower) inclination toward the advantageous, by which the human agent can desire and pursue only what is good for him, and the (higher) inclination toward justice, in virtue of which he is capable of wanting and seeking what is good in itself. Scotus holds that the will's 'innate freedom' (to be explained later) lies in its affection for justice (Scotus, 1986, p. 153), so that this freedom may be understood as consisting in the human agent's capacity to rein in, or transcend, his affection for the good-for-me. For Scotus, 'the existence of [these inclinations] in the will is proved by the conflict that one may experience within himself, as in the case of martyrs who choose to die by a free act of their will despite the natural fear they have for death' (Bonansea, 1965, p. 87).

So the human will is a self-determining power for opposites and it has two basic desires. But what is the relation between these properties of the will? This has puzzled interpreters. Two competing interpretations of this relation - Boler's (1993) and Lee's (1998) - are worth considering. According to Lee's (1998) interpretation, the will is a self-determining power for opposites in virtue of having the dual affections (and of being able to choose between them). So the dual affections are both necessary and sufficient for the human will to be this sort of power. On this interpretation, the innate freedom of the will consists exclusively in the possibility of restraining the lower affection by the higher (nobler) affection. This implies that human freedom just is moral freedom: 'The freedom of human agents is [...] the ability to transcend [...] the desire for one's advantage' (Lee, 1998, p. 52).

This interpretation ascribes to Scotus a Kantian conception of free will, with all its implications (especially for moral imputability), such as: (1) actions motivated by the affection for personal advantage cannot be free, and their authors morally accountable. That is, immoral actions are not free, and so not morally imputable. (2) All actions triggered by the affection for justice are free, and their authors morally responsible, no matter how they happen to originate from this affection. Think of the actions of the Simple Moral Wanton and of the OCPD-type Moral Wanton. But these damaging implications cannot be a reason to reject Lee's interpretation of Scotus's account of the nature and freedom of the will. It can reasonably be rejected only if it is not consistent with Scotus's doctrine.

I think that Lee's reading is indeed inconsistent with Scotus's voluntarism, which stresses the broad scope of human freedom. In support of this I will cite just two pieces of textual evidence. The first piece is Scotus's statement that 'it is in the will's power not only to will this way or that way but also to will or not to will, since the will has freedom with respect to acting or not acting' (Scotus 2017, *Ordinatio* 1, d. 1, part 2, q. 2, paragraph 149). If the will is free to act or not to act, this freedom necessarily extends to actions motivated by the desire for the advantageous. The second piece is Scotus's account of one particular kind of deliberate sinning: 'the will chooses for itself to will evil in virtue of its sheer freedom and without any motive or explanation external to itself' such as a 'passion in the sensory appetite or error in reason' (*Ordinatio* II, d. 43, q. un., paragraph 6). This exercise of one's freedom is inconsistent with the view that human freedom just is moral freedom. And so Lee's interpretation must be rejected.

Now according to Boler's (1993) interpretation, the will's capacity for self-determination between opposites is a capacity to will or not to will to act (irrespective of the antecedent conditions): when the will freely chooses or refuses to act, it must, at the time of choosing or refusing to act, be capable of the opposite. Boler calls this the 'metaphysical' freedom of the will (pp. 112-3). On this interpretation, the possession of the dual affections (for the advantageous and for justice) is neither necessary nor sufficient in order for the will to be a self-determining power for opposites or to have and exercise metaphysical freedom. As to what Scotus calls the 'innate freedom' of the will and associates with the affection for justice, Boler takes it to be exclusively a moral phenomenon, one that perhaps presupposes metaphysical freedom but is independent of it. The will's 'innate freedom' is a moral phenomenon because it consists in the

ability to choose to rein in or not to rein in our affection for what is advantageous to us for the sake of what is just. So, the dual affections of the will allow for the sort of complex motivational profile that moral life requires. They ‘provide the basic parameters for what it makes sense to say about moral action: morally good or evil action is succeeding or failing in transcending one’s natural inclinations (p. 125).

This interpretation of Scotus’s account of the nature and freedom of the human will seems very consistent with his voluntarism. If we take it to be Scotus’s own conception of the (free) will, then we can easily see how it would be better suited for Kant’s theoretical purposes, namely, securing genuine self-determination in action, securing the rational necessity of morality, and securing the motivational appeal of morality. First, by distinguishing between metaphysical freedom and moral freedom, with the former as a pre-condition of free agency, this conception secures the possibility of genuine self-determining agency. On this conception, both morally good and morally bad actions can be free, self-determined actions. Second, by treating the affection for justice (a moral desire) as a basic inclination of the will, this conception suggests that morality (or the moral law) is a volitional necessity (rather than a rational necessity, for Kant). And, third, it easily secures, at the same time, the motivational force of morality (or the moral law). What is striking here is that Scotus’s concept of the free will achieves all these feats, where Kant’s own concept was in trouble, while it is welded to the same view of morality as Kant’s: that the moral life is not one of the pursuit of personal happiness but one of commitment to transcending our inclination toward our happiness, and that in this commitment lies our higher vocation or (moral) freedom. So re-naturalising the will, or bringing self-determination and motivation back within nature, holds the promise of saving both freedom and morality.

Chapter Five: Action, Reason and Will: Consciousness, Judgement, Commitment

Human beings are not alone in having desires and motives, or in making choices. They share these things with members of certain other species, some of which even appear to engage in deliberation and to make decisions based on prior thought. It seems to be peculiarly characteristic of humans, however, that they are able to form [...] “second-order desires” [...]. No animal other than man [...] appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires (Frankfurt, 1988, p. 12).

‘There is a sense in which it may be said that each of us lives in a world of his own,’ notes Lonergan. ‘That world usually is a bounded world, and its boundary is fixed by the range of our interests and our knowledge’ (2016, p. 60). If the reach of our knowledge and interests sets the boundaries of our world, it also constrains our actions. In fact, what we do crucially depends on what we think and what we want, or on our reason and our will. By ‘our’ I mean to emphasise the powers of reason and will specific to the sort of creatures that we are. Presumably, we are not the only species endowed with reason and will. The difference between a dog’s powers of reason and will and a human being’s such powers mirrors the difference between the sort of creatures that dogs and human beings are. So understanding human action - that is, intentional action by a human being - requires grasping the sort of creature that a human being is, on the one hand, and the nature of reason and will and the relation between them, on the other. But both the nature of the human being and the nature and relation of reason and will are chronically perplexing and contentious.

In the previous three chapters, I examined three different accounts of the constitution of human action developed against the backdrop of the widespread dissatisfaction with the desire-belief theory of action. According to this theory, when we act intentionally, we perform the action in the belief that it is a means to the fulfilment of a desire we happen to have. And the desire-belief pair rationalises the said action. The main trouble with this theory is that the agent it portrays is not recognisably human. The desire-belief agent is not a full-blown human agent because he lacks the capacity for a reflective evaluation of his desires and the capacity for self-

determination in action. This agent's powers of reason and will seem to be unlike those of fully developed human beings.

The three alternative accounts of the constitution of human action that I discussed in this thesis are by Bratman, Velleman and Korsgaard, and I found all of them wanting in some respects. I argued that Bratman's account is wanting due to its flawed theory of the nature of the will or intention. Bratman holds that we are planning agents - creatures that plan for the future and that need plans to coordinate with themselves and with others. These plans are embodied in our intentions to act, and they are subject to the structural and instrumental principles of practical reason, namely, the norms of attitudinal consistency and means-end coherence. I argued, *inter alia*, that some of our intentions, namely, proximal intentions, do not seem to fulfil any coordination function, and that we often act intentionally without any intentions or plans thus construed (- as when we act spontaneously).

As regards Velleman's account, I raised some serious doubts about its premise that the special way in which we know what we are doing is the gateway to understanding human agency. Velleman claims that we are self-knowing agents, or desirers of self-knowledge, and that intention is a cognitive attitude by which we commit ourselves to do what we are prepared to know and understand. I argued that this notion of intention and the claim about the distinctiveness of our practical knowledge are both implausible.

Finally, about Korsgaard's account I argued that the moral law as the law of volition and agency cannot be logically derived from the assumption of the deliberating agent's freedom from his natural inclinations. According to Korsgaard, a deliberating agent should conduct himself as if he were transcendently free, and this entails choosing in accordance with the categorical imperative. I have disputed this entailment, showing that it has some perversely counterintuitive implications for autonomy.

In this penultimate chapter, and the final one, I attempt to develop a novel account of the constitution of human action by tackling the issues of the nature of the human being and of the nature and relation of reason and will. This chapter will address the issues of the nature of the human being (mainly through an analysis of the structure of human consciousness), and of the

nature of reason and will. And the last will deal with the relation between reason and will in the constitution of human action.

About my account of the constitution of human action two clarifications deserve to be made at this point. The first clarification is that this account is a continuation of the view of human action as intentional action by a self-determining creature that I proposed in the first chapter. Recall that I argued that intentional action is action done for reasons and that intention is directly connected to the action it makes intentional. I also argued that intentional action is action done for reasons because human action is fundamentally and distinctively purposive. The purposes that drive our various intentional actions are the reasons for which these actions are performed.

The second clarification is that, like the alternative accounts considered above, I begin my own account with the desire-belief model (- which is perhaps the most basic account of the way in which what we do depends on what we think and want -) as a theoretical foil and point of departure. I argued in the first chapter (subsection 3.3) that this model does not capture human agency. By implication, the desire-belief agent is not a truly human agent. He lacks a range of psychological capacities that constitute us as human agents. I don't think that what the desire-belief agent needs in order to become a truly human agent is Bratman's plans, Velleman's drive for self-intelligibility, or Korsgaard's transcendental freedom and respect for the moral law. I believe that all he requires is the capacity for human consciousness, practical reason as the capacity for evaluative thinking about what to do, and the will as the affective capacity for intention and commitment.

In what follows, I develop accounts of human consciousness, practical reason, and the will with a view to explaining the constitution of human action. I first argue that human consciousness is distinctively structured by an enquiry-guided pattern of four levels of conscious intentionality. The basic idea here is that our consciousness is structured in such a way that we're characteristically inquisitive about, and hence active with respect to, our inner and outer experiences. And this internal structure is critical for distinctively human agency because it makes possible the exercise of practical reason and the will (Section 1). I then develop an account of practical reason as our capacity for evaluative thinking about what to do. I argue that practical reasoning concludes in judgements of value about what to do, and that these

judgements are essentially constituted by feelings of value, factual judgements and imaginative judgements (Section 2). Next, I sketch an account of the will as a distinctive capacity of the mind, and argue that the commitment involved in a state of intention is an existential commitment (Section 3). Finally, I briefly conclude and sum up this chapter (Section 4).

1

Let me begin with human consciousness. I propose here that to understand the nature of the human being it is necessary to understand the structure of human consciousness. Consciousness is probably the basis of every (distinctive) human experience. And many philosophers see consciousness as critical for understanding agency, autonomy and even normativity. In this section, I present a brief analysis of the structure of human consciousness, which will form the basis of my discussion of the nature of practical reason and the will in the rest of this chapter, and of the relation between them in the next. This analysis is Lonergan's own analysis (or, at any rate, my grasp of it), which I started in the first chapter (subsection 2.2) in response to Hume's bundle theory of the self.

As a recall, I argued there that human consciousness has a bimodal character. In one mode, consciousness is awareness of objects and their felt qualities through mental operations (such as seeing, hearing, touching, smelling, tasting, imagining, etc.). In another mode, consciousness is a concomitant awareness of oneself as the subject of the awareness of objects and their properties. That is, in the very act of perceiving an object I become aware not only of the object but also of the one (myself) perceiving and the act of perceiving. I called these two modes of consciousness *intentionality* and *self-consciousness*.

But this is not all there is to human consciousness. If it was, my claim (made above) that the desire-belief agent lacks the capacity for human consciousness would certainly be false. The desire-belief agent's consciousness certainly has this bimodal character. The fact that this agent is able to form beliefs about how to satisfy his desires is proof that he is aware of his desires and their objects and that he is aware of himself as having them. So my claim about the consciousness of the desire-belief agent should be understood as recognising its bimodal

character and as suggesting that it misses something else in order to be a human consciousness, something presumably distinctively human. Now what could this additional ‘something’ be?

Many theorists in philosophy, psychology and biology seem to identify self-consciousness as a distinctive trait of human consciousness, but they do not have a univocal concept of it. By self-consciousness, some mean awareness of one’s own mental states (introspective awareness), and others awareness of oneself as the subject of one’s conscious experiences (self-presence). They claim, in other words, that humans can, but lower animals cannot, be aware of their own mental states or aware of themselves as attending to objects in the world. I think that these claims are implausible. Studies have shown evidence of metacognition, one of the components of introspective awareness, in a range of lower animals. When faced with difficult tasks in which there is some degree of uncertainty regarding the outcome, dolphins, rhesus monkeys and rats - like humans - have been shown to choose easier tasks with higher chances of a successful outcome. This lack of confidence in their ability to perform a particular task coupled with the corresponding choice of an easier one is a potential marker of metacognition (Lage et.al., 2002).

As regards self-consciousness as self-presence, the fact that non-human animals have an episodic-like memory capability suggests that they have a certain form of conscious experience of themselves as doing, or attending to, things. Potential evidence of this capacity in scout bees, for instance, is that they can recall and communicate to their swarms, through the waggle dance, the distance and direction of suitable places for a hive that they have found during their search (Hyman, 2015, p. 47). The fact of being conscious while attending - that is, the fact of being self-present or self-conscious - is what makes it possible for us to report our experiences. For example, in order for me to be able to report to someone that ‘I was listening to a Michael Jackson song some minutes earlier,’ I must have been aware of myself as the subject of this listening experience during its unfolding. During the listening experience I was attending to the song, and not to myself. Now to be able to report this activity to which I was not attending, I must have had some awareness of my(self) listening even while attending to the song itself (Helminiak 2021: 1). The same might hold true for scout bees insofar as they are able to report their past conscious experiences. They must have been aware of themselves as subjects of those experiences. If so, there’s no reason to think that humans are unique in being introspectively aware or self-present in their conscious experiences.

Now contrary to these theorists, Korsgaard (2014) grants self-consciousness to lower animals but holds that human self-consciousness is distinctive in that we humans, but not brutes, can be aware, when experiencing objects, that our minds' activity is shaping the character of our experiences. I think this claim is simply false. That the workings of our minds are influencing how we experience the world is not something that can be given to us in consciousness. When you are looking at the glass of water on the table in front of you, you cannot experience that the activity of your mind is determining what you perceive and how you perceive it. This is something we learn from science, and not a datum of our conscious experiences. As awareness of the influential activity of one's mind is not at all an experience that one can have, it just cannot be a trait of human consciousness, let alone a distinctive one. In sum, it is not true that the capacity to be self-conscious, as construed by various theorists, is the fundamental difference between humans and other species. Humans do not seem to be unique in having both conscious and self-conscious experiences.

What, then, needs to be added to the bimodal character of the desire-belief agent's consciousness in order for him to have a human consciousness? According to Lonergan¹, the structure of human consciousness is not only bimodal but also four-levelled and enquiry-guided. To see this, imagine that you're reading the *Guardian* and suddenly your attention is arrested by a political cartoon². You look at the funny figures, burst into laughter, read the caption, and look back at the cartoon, and wonder: 'what does it mean?' You keep looking and looking, assuming that there is in the cartoon a clue that you are failing to grasp. Finally, after several minutes of puzzling, you get the point (or so you believe). But you may not stop there. After you've got an insight into the cartoon before your eyes, you might wonder whether this insight is right - whether, say, the cartoon is saying that the Tories have no solution to the cost of living crisis. And, having determined whether your insight is correct, you may wonder, finally, what to do in light of your

¹ My discussion of Lonergan's analysis of the structure of human consciousness draws mainly on *Insight: A Study of Human Understanding* (1992 [1957]), *Method in Theology* (2017 [1972]) and *The Ontological and Psychological Constitution of Christ* (2002). I have chosen not to mention text and page references in much of the discussion because of this multiple textual origin of the ideas examined.

² This example is adapted from an example by Flanagan in his 1997 book *Quest for Self-Knowledge: An Essay in Lonergan's Philosophy*, p. 16.

judgement or reflective understanding of the cartoon. Will you, for example, act on this judgement by, say, voting against the Tories in the next general election?

Your experience with the cartoon illustrates the additional (and presumably distinctive) features of the structure of human consciousness. According to Lonergan, human consciousness is structured by a pattern of four successive levels of conscious intentionality³: experience, understanding, judgement and decision. By experience, Lonergan means ‘a preliminary unstructured sort of awareness’ (2002, p. 157) of something. The awareness is ‘preliminary in that it is a prerequisite to intellectual inquiry’ (p. 159) or to the activity of understanding, and it is unstructured in that it still has the form of raw material. So, as conscious human creatures, we experience things in the world (through our senses) and things in ourselves (such as sensations, feelings, desires, memories, etc.); we strive to understand our experiences of those things; we form judgements of our understanding of them; and we often decide courses of action based on those judgements. But we, human beings, are most certainly not alone in experiencing and understanding the world and ourselves or in forming judgements and making decisions. We share these mental activities and the underlying capacities with other species.

What is presumably unique to us is the patterned and enquiry-guided working of our minds. As inquisitive and language animals, we are not at all content with just having sense or mental experiences. We typically question them. Asking a question - what, why, how or what for? - about our experience of data spurs the upward movement from the empirical level to the intellectual (understanding) level. The insight gained at this level answers the question raised by the data, but it may also raise a question about its own adequacy, correctness or truth: Is it true? Is it really so? This question for reflection leads to the rational (judgement) level, where we either affirm or deny that our insight makes sense of the experienced data. And this judgement may prompt a question regarding what to do about it, thereby summoning the volitional (decision) level of conscious intentionality. Thus, we inquisitively move from experiencing to understanding to judging to deciding. The desire-belief agent is not a truly human agent because

³ The nominal phrase ‘conscious intentionality’ is meant to convey the bimodal character of human consciousness, as explained above. That is, whenever we perform any of the operations on each of the four levels, we are simultaneously aware of the objects we are attending to and of ourselves as the operating subjects.

his consciousness does not seem to have this inbuilt, transcending pattern of attention and enquiry.

This unique pattern of our mind shapes our relation both to the external world and to our own mental states. As such, it is critical for rational agency. It enables us to take ourselves seriously in our thinking and our acting - that is, to ensure that our perceptions, our thoughts, our feelings, our choices and our behaviours withstand reflective scrutiny (Frankfurt, 2006, p. 2). It makes it possible for us to act on reasons, rather than being simply moved to act, by enabling us to step back from the immediacy of our experiences and motives and to develop reflective higher-order responses to them. In a nutshell, this internal device of our consciousness is critical for distinctively human agency because it makes possible the exercise of practical reason and the will.

2

Let me take up practical reason here. I've just argued that the distinctive inbuilt pattern of our minds makes it possible for us to act in light of our conscious appreciation of the reasons we have to act. Because we are inquisitively self-conscious creatures, we do not take each of the desires in our psychic economy as a reason for action, unlike, presumably, the simple desire-belief agent, who treats each of them at least as a *pro tanto* reason. Instead, we regard the experience of desires as an opportunity to take a reflective stance on them and determine whether or not they are worthy of endorsement and satisfaction. And we do so by exercising our practical reason.

But what is practical reason? In its most basic form, practical reason is perhaps just the capacity for determining, through some kind of reasoning, what to do in one's situation. Thus understood, this capacity can plausibly be found amongst members of non-human species. And the desire-belief agent certainly has this capacity. He can be said to reason about what to do given the desires he has, especially when they conflict. But given the complex structure of consciousness within which it is embedded, human practical reason is much richer. It is our capacity for determining, through a more robust kind of reasoning, what we ought to do or what we have

(most) reason to do. This reasoning concludes in a belief or judgement about what to do given the reasons that we take ourselves to have.

In this section, I want to argue that practical reason is our capacity for *evaluative* reasoning about what to do, and that the judgement in which practical reasoning issues is essentially a judgement of value. This argument will run as follows:

P1. Practical reasoning issues in a judgement about what one has most reason to do.

P2. A reason for an agent to A makes A-ing intelligible to him (in a certain sense of 'intelligible').

P3. For an action done for a reason to be intelligible (in the relevant sense) to an agent just is for it to be seen as good (in a certain sense of 'seen as good').

C1. If a reason for an action makes that action intelligible, and if for an action done for a reason to be intelligible just is for it to be seen as good, then a judgement about what one has most reason to do is a judgement of value (or about the goodness of what to do).

C2. And if a judgement about what one has most reason to do is a judgement of value, and if such a judgement is the issue of the activity of practical reasoning, then practical reason is a capacity for evaluative reasoning about what to do.

I will proceed as follows. I will provide an extended argument for P1, - that is, for a narrow view of practical reasoning as ending in a judgement (against the broader view that it concludes in an intention or even in an action) -, in Chapter 6, sub-section 3.1. I will not argue for P2, which is the widely accepted explanatory constraint on reasons for action. That is, reasons must be the sort of things that explain, or make intelligible, the actions for which they are reasons. When you act for a reason, the reason explains why you so act, or it shows, at least from your point of view, why it makes sense to act as you do. I will instead argue that the dichotomy between the agent-relative and the agent-neutral senses of intelligibility is false. As concerns P3, which is the Guise of the Good thesis, I will assume its truth and then just proceed to reconcile, or dispel the dichotomy between, the two broad families of accounts of the appearance of the good. Finally, regarding C1 and C2, they do not need arguing for, as their truth value is logically dependent on that of premises P1 through P3. In proceeding this way, I will be arguing that practical reason is our capacity for evaluative reasoning about what to do (2.1) and that practical reasoning as

evaluative thinking (that concludes in a value judgement about what to do) also centrally involves imaginative thinking (2.2).

2.1. Now assuming, at this point, the truth of the view that practical reasoning ends in a judgement of what one has most reason to do (P1), I will plunge straight into the enquiry of the sense in which a reason to A makes A-ing intelligible to its agent. The first thing to note in this connection is that the reason that makes an action intelligible to its agent is the reason in light of which he acts or the reason that furnishes his motivation. It is, in other words, the reason that would constitute the agent's own (positive) answer to an Anscombean why-question.

As I am writing this section of the chapter, I feel that my ability to focus has dwindled. I stand up, grab my winter coat, and go out for a walk. Why do I go out for a walk? Different stories can be told to answer this question and make my action intelligible. One story might hold that my getting up and going out for a walk is all due to the complex but random interactions between neurones in my brain. Another story might explain that I am going out for a walk out of a strong, repressed desire not to walk along the professional path set out for me by my domineering academic father. Yet another story, this time by someone more familiar with my daily routine, could be that I am going out because my doctor has instructed me to exercise every day for at least 20 minutes.

All these stories might be very plausible explanations of my action of going out for a walk. And yet they might not be valid answers to the why-question addressed to me, or they might fail to capture the reason(s) in light of which I act as I do. In so failing, they will not make my action intelligible to me as its author. To explain my action in a way that makes it intelligible to me as the agent, it seems that any story told about it needs to capture the circumstances of my action as they appeared to me or as I represented them to myself (- I was writing this part of the chapter, I felt I had become unable to focus, I thought that a walk would replenish my power for concentration, etc.). And in capturing these circumstances, the story about my action will explain it and show how it made sense to me even if my representation of them involved an error of factual or normative judgement. It could be the case, for example, that exposure to the cold during the walk will trigger a migraine attack and so make it impossible for me to continue working on this chapter, or that what I actually should do to regain my ability to focus is take a

nap. But, in spite of my erroneous representation of them, the circumstances of my action and the considerations in light of which I act are what explains my action and makes it intelligible to me as the agent.

It would seem to follow from the foregoing that the intelligibility of an action done for a reason is agent-relative rather than agent-neutral. But this appearance is deceptive. In fact, I want to reject the dichotomy between agent-relative and agent-neutral accounts of the intelligibility of action. According to agent-relative accounts, an action done for a reason is intelligible to its agent only in virtue of its relation to some consideration that the agent has in mind, generally a property the agent takes to be instantiated by it (Boswell, 2018, p. 8). But proponents of these views recognise that not just any mental consideration will count as an intelligible-making property of action. That is, actions are not always intelligible in the light of the mental considerations that lead to them. For example, Quinn's Radio Man's mere urge to turn on any radio that he happens to see (Quinn, 1993, p. 32), or Clark's shoe-painter's intrinsic desire to spray paint on his left shoe just to make it weigh a little more than his right shoe (Clark, 2010, pp. 234-5). Relating the action of turning on radios to the Radio Man's compulsion, or the action of painting the left shoe to the painter's inclination, will fail to make them intelligible. These issues have led agent-relativists about the intelligibility of action to acknowledge that, though intelligible-making properties are subjective, they must be subject to some kind of substantive constraints. By contrast, according to agent-neutral accounts, intelligibility is an objective property of actions; it is not in the eye of the beholder (MacIntyre, 1986, p. 64) much like, say, the intelligibility of a sentence. For example, the unintelligibility of Chomsky's utterance 'Colourless green ideas sleep furiously' is not in its utterer's eye or mind. On these accounts, the action must be intelligible to, presumably, the members of the rational community the agent belongs to.

I think that the dichotomy between agent-relativity and agent-neutrality about the intelligibility of action arises out of the neglect of the nature of human agents. By nature, we are social creatures, and not Leibniz's windowless monads. We are social in the sense that we live with others, we need each other, and we seek and enjoy communion, love or friendship. But, more fundamentally, we are social in the sense that interaction with others is a necessary condition of our existence and flourishing (Gawande, 2009). Our Sociality, then, is not a contingent fact

about our human nature. Now it seems intuitive that practical reason, like any constitutive capacity of an agent, can only be grounded in the nature of the agent whose capacity it is. That is, an agent can exercise his practical reason only as the sort of creature that he is, or only in conformity with his own nature. If so, then we, human beings, exercise our practical reason as social creatures.

It is indeed undeniable that we engage in practical reasoning as social creatures. In thinking about what to do, we are typically aware that we act in the social world, where our actions have consequences on our lives and the lives of others, where most of the important goals we have depend on a complex network of social practices, and where their successful pursuit requires coordinating with other people. And this is not just some sort of idle awareness. Our social awareness is a determining feature of our practical thought, motivation and behaviour. It constraints us, amongst other things, to take ourselves seriously as social creatures. Philosophers have written eloquently and convincingly about our deeply ingrained impulse to take ourselves seriously as rational creatures (Korsgaard, 1996; Frankfurt, 2006; Velleman, 2009), but they have systematically overlooked part of the picture of the motivation underlying this natural tendency or disposition: it is our social desire to be intelligible to ourselves and others that puts us under pressure to ensure that our actions make sense or withstand reflective scrutiny. Taking ourselves seriously as social agents, or factoring our social nature and circumstances into our practical reasoning, means that no consideration will count as a reason for action to us unless it can make the action it supports intelligible to us and to our kind, or in our social environment.

So when we say that a reason for an agent to A makes A-ing intelligible to him (P2), it is wrong to add the qualification “in a certain sense of ‘intelligible’” if we mean to distinguish between the agent-relative sense and the agent-neutral sense, and if the relevant agent is a human agent. The dichotomy between agent-relativity and agent-neutrality about the intelligibility of action seems incompatible with facts about the social nature of the human agent. As fundamentally social creatures, our ‘actions pass through the private world of the mental as well as the public world of the social’ (MacIntyre, 1986, p. 65). On the one hand, we are self-aware, rational agents whose actions are guided by our individual appreciation of what reasons we have to act. On the other hand, most of our reasons for action depend on the existence of social practices and most of our actions require interaction with others. Because of the social preconditions of human

agency, the relevant sense of intelligibility of an action to its agent is intelligibility to him as a social creature. And this sort of intelligibility does not admit of the dichotomy between agent-relativity and agent-neutrality.

Now for an action done for a reason to be intelligible to its agent just is for it to be seen by him as good - in a certain sense of 'seen as good' (P3). This claim views the Guise of the Good thesis as the (only plausible) satisfaction condition of the intelligibility or explanatory constraint on reasons for action. According to one version of this thesis, when we act intentionally, or for a reason, we see that action as good⁴. It is controversial amongst the proponents of this thesis what sort of mental state or attitude is this 'appearance' of the good. And the thesis itself, which is an ancient doctrine in the philosophy of action and metaethics, has recently generated a controversy between its critics (e.g. Stocker, 1979; Velleman, 2000; Setiya, 2007) and its defenders (Raz, 1999; Orsi, 2015; Saemi, 2014 & 2017; Boswell, 2018). And so the claim that for an action to be intelligible to its agent it must be seen by him as having a desirability characteristic or as instantiating an evaluative property must be controversial and hence in need of a defence. But, as I announced earlier, I will not argue for the Guise of the Good thesis here. I will simply propose that the social nature of the human agent might lend support to this thesis. The point here is that there seems to be no explanatory device that better makes an action intelligible interpersonally or in its social context, or that better escapes the threat of infinite regress, than an appeal to its apparent instrumental or final value. Given our sociality, then, the intelligibility of our actions is probably best understood as a property that our actions have in virtue of their relationship to the goods they are perceived to instantiate.

Putting aside the controversy about the plausibility of the Guise of the Good thesis, I will now turn to the controversy about the nature of the state of 'seeing as good' in an attempt to reconcile, or dispel the dichotomy between, the two broad families of accounts of the appearance of the good. According to one family, the state of 'seeing' something under the guise of the good is a desiderative state (a desire or a desire-like attitude), whereas, according to another, the appearance of the good is a doxastic state (a belief or a belief-like attitude). The dichotomy between these two families of accounts of the Guise of the Good thesis arises mainly out of a

⁴ Other versions target pro-attitudes like desire and intention, and hold that whatever is desired or intended is desired or intended under the guise of the good.

disagreement about the epistemology of value - that is, about how we grasp or access value - and out of each family's commitment to the view that only one aspect of our mind or psychology is implicated in our conscious experiences of value. So if it is the case that there is not only one value-attuned state of consciousness, or only one guise of the good, then the dichotomy will cease to exist.

It is more plausible, I think, to speak of two different but related guises of the good - the affective guise and the cognitive guise - and to hold that the affective guise is the primary guise and the cognitive guise is the secondary guise, and that the affective guise is the basis of the cognitive guise. That is, a fuller human grasp of value involves both affectivity and cognition, moving from the former to the latter. But what does each of these guises contribute to the fuller grasp of value? We have first a direct perception or intuition of value and disvalue in and through our feelings or affective experiences. Think of our experiences of love and hate, of pleasure and pain, of joy and sadness, and of excitement and anxiety. Through the feelings or emotions that things arouse in us we directly grasp their goodness or badness. And these affective apprehensions of value and disvalue are grounded solely in the heart's own reasons that are perhaps unknown to reason. As such, they are not responses to a cognitive grasp of the nature and qualities of objects, events or states of affairs (Steenburg, 1994, p. 203). Instead, they form the basis for cognitive apprehensions or judgements of value. So value judgements as conclusions of practical reasoning depend on affective perceptions of value, or include these perceptions as their constitutive components. But the affective and cognitive guises of the good can and often do go apart, as evidenced by the form of *akrasia* that consists in acting against one's best judgement.

2.2. Having established, hopefully, that the judgements of practical reason are value judgements about what to do, I want to devote this sub-section to a discussion of other components of these judgements. Many philosophers who identify feelings of value as the primary component of value judgements typically recognise just one other element: factual thinking or judgement (e.g. Hume, 1998 [1777]; Lonergan, 2017 [1972]). The basic idea about the judgement of fact is that it involves the knowledge of one's individual psychological characteristics (such as interests, purposes, capacities and limitations) and the knowledge of the appropriate means, the social

context and the possible consequences of one's actions, which are necessary for rational agency. And that is why factual judgement is important for evaluative judgements of what to do.

But I think that there is another important component of practical reason and value judgements that has been heretofore overlooked. This component is the imagination or imaginative reasoning. In my view, practical reason also centrally involves the imagination. *Practical reason centrally involves the imagination because practical reasoning is bound up with self-projection and imaginative reasoning involves imagining oneself in a future or another self* (about which more shortly). This claim about imaginative reasoning as another central component of practical reasoning raises at least two questions. The first is this: since imagining is a very common mental activity that comes in various forms, what type of imagining, if any, is an element of practical reasoning? And the second is the question of how the relevant variety of imagining can be epistemically useful as a constituent of practical reasoning. These questions are very complex and giving them a proper treatment is outside the scope of this section. In what follows, I first identify and describe the variety of imagining that is relevant to practical reasoning. And then I briefly explain how it can be used in practical reasoning for epistemic purposes and benefits, and how we are active in employing it.

Let me begin with the deliberatively relevant case of imagining. A proper description of this case requires distinguishing it from other cases. And for this task of surveying the varieties of imagining I will draw on an illuminating discussion by New (1999). Now New distinguishes several different cases of imagining, of which only the following three are relevant for my purpose here. First, there are cases of imagining that are cases of supposing, assuming or being inclined to believe (p. 69), as in Jack's answer 'I *imagine* so. I was there about a minute ago' to Sally's question 'Is the convenience store on the street still open?'. Second, some cases of imagining are cases of having a false belief or a delusion (p. 69). An instance of this is when, for example, a nurse at a psychiatric hospital reports about a patient that 'he *imagines* he is Jesus the Son of God,' or when we say that the crabs that a hallucinating Sartre saw following him in the streets of Paris after he was injected with mescaline were only 'imaginary'. And, third, there is a type of imagining that involves, or is essentially about, 'thinking of the probable consequences of a proposition hypothetically supposed (imagined) to be true' (p. 70). It is in this sense that 'imagine' is used when someone asks you to 'imagine what the political situation of Russia

might have been today if Trotsky, rather than Stalin, had succeeded Lenin'. This type of imagining, unlike the others, is a kind of reasoning. It 'qualifies as imagining because the reasoner does not believe the consequences did happen, but only that they might or would have happened if the supposition made had been true' (p. 70). And it qualifies as reasoning because it involves thinking logically or drawing logical inferences. Now it is this form of imagining that is important for practical reasoning.

Practical reasoning, I have argued, issues in an evaluative judgement about what to do. I am now proposing that it also centrally involves imaginative reasoning. This proposal immediately raises the question of how imaginative reasoning is distinctively involved in practical reasoning and how it distinctively contributes to the value judgements of practical reason. To see this, we need first to appreciate that acting is bringing about desired or intended change in ourselves and in the world, and so changing ourselves and the world in a desired or intended way. This, I take it, is pretty much uncontroversial and does not need arguing for. Now imaginative thinking might distinctively support practical reasoning about what one rationally ought to do in at least two related ways. The two ways explain the claim (stated earlier) that *practical reason centrally involves the imagination because practical reasoning is bound up with self-projection and imaginative reasoning involves imagining oneself in a future or another self*. The two ways are:

First, insofar as acting is changing ourselves in a desired or intended way, reasoning about what one ought to do needs to involve picturing and evaluating the future self that one desires or intends to be. And this is where imaginative reasoning comes in. In imaginative reasoning we *put ourselves in the shoes of our projected or future selves* and gain a distinctive kind of understanding of ourselves and our expectations. Let me illustrate this distinctive feature of imaginative reasoning by drawing a contrast between factual reasoning and imaginative reasoning as constitutive elements of practical reasoning and judgements of value. Consider a case of reasoning about whether to be an actor or a lawyer, whether to live in the city or in the country, or whether to have children of one's own or to adopt. In this case, the contribution of factual reasoning might consist in toting up the pros and cons of each option. By contrast, the contribution of imaginative reasoning will involve imagining oneself in each of the positions - as an actor, as a lawyer, as a biological parent, as an adoptive parent, etc. - that is, imagining various projected selves, feeling what it feels like to be each of these future selves and, through

those feelings, apprehending the value or disvalue of being these selves. It follows that imaginative reasoning distinctively contributes, *inter alia*, feelings of value, which, as I argued earlier (2.1), are the primary data of value judgements about what to do.

Second, and relatedly, insofar as acting is also changing the world in a desired or intended manner, and insofar as we engage in practical reasoning as social creatures, reasoning about what one ought to do needs to involve picturing and evaluating the impacts of one's projected actions on others or the world. Imaginative reasoning may enable us to do exactly this. For, by way of the imagination, we can (and often do) *put ourselves in the shoes of other selves* or other people, and gain an understanding of how they might feel about our actions and their effects on them, and envision their reactive attitudes toward us. Imaginative thinking thus involves taking an empathetic stance. And on the basis of the insights gained from taking this stance, we may (and often do) judge what we ought to do.

Imaginative reasoning, then, plays a fundamental role in practical reasoning as evaluative thinking. It distinctively contributes to the formation of value judgements about what to do in that it involves thinking about the probable consequences of states of affairs supposed to be true and it enables the agent to put himself in the shoes of his future self or another self. But in order for a deliberating agent's imagining to be of epistemic value to him, he must appropriately control both its initial content and its unfolding process (Badura and Kind, 2021, p. 2). Now we can take active control of our imagining by 'deploying it in the consideration of our situation, and also by [questioning] how well our imagination has worked and taking steps to employ it better, for instance, by paying more attention, or setting aside the distortions of prejudice or fantasy' (Stern, 2022, pp. 15-16), etc. Thus actively controlled by its agent, an episode of imaginative reasoning can provide, along with affective value-perceptions and factual thinking, valid grounds for practical inferences or evaluative judgements about what to do.

3

Now practical reasoning, as evaluative reasoning that concludes in a value judgement about what to do, is a crucial part of the exercise of the will or the formation of intentions. This ingredient of a distinctively human agency is missing in the desire-belief model. In fact,

Davidson, the modern father of this model, later came to disown it for this very reason. That is, he realised that the model did not have the psychological resources to explain the mental state of intention, which he thought was a necessary component of a theory of intentional action. Though a welcome addition to the desire-belief model, some philosophers thought that the introduction of an intention was not sufficient. They argued for ‘the need to return to a substantive conception of the will as a distinctive faculty and to allow for the presence of *volitions*’ (Carey, 2017, p. 71). But while volitions as *sui generis* mental states are still frowned upon in analytic philosophy of action, presumably because of their lingering whiff of mystery, intention has been canonised as a distinctive action-explaining attitude.

In this section, I want to defend a substantive conception of the will as a distinctive capacity and as an affective capacity for intention. This view of the will is unique in the contemporary action-theoretical literature in at least three respects. First, it is a synthesis of two rival concepts of will captured in here:

If we consider that a person’s will is that by which he moves himself, then what he cares about is far more germane to the character of his will than the decisions or choices he makes. The latter may pertain to what he *intends* to be his will, but not necessarily to what his will truly *is* (Frankfurt, 1988, p. 84).

Implicit in this statement are two different concepts of the will in the literature. The first is the concept of the will as the capacity to determine consciously and independently (of one’s tendencies, inclinations or dispositions) what to do (Hieronymi, 2022, p. 13). The second is the concept of the will as ‘that collection of more-or-less ordinary, interacting aspects of the person’s psychology (their cares, concerns, beliefs, desires, commitments, fears, etc.) that generates intentional, voluntary, or responsible activity’ (Hieronymi, p. 14). I call the first view the *decisionist concept of will* and the second the *desiderative concept of will*.

By conceiving of the will as an affective capacity for intention (or choice or decision) I obviously blend these two traditions. But this is not a fortuitous or capricious theoretical gesture, or a strictly strategic manoeuvre. There are basic facts about us that warrant it. Let me begin with facts about the ‘affective’ (or desiderative) side of our faculty of volition. One such fact, or basic intuition, is that human beings are inherently appetitive creatures. Some of our appetites or

inclinations are innate to, or grounded in, our nature and define us as the sort of creatures that we are. Consider, for instance, our natural inclinations toward such intrinsic goods as self-preservation, mental and physical integrity and health, friendship and communal life, self-fulfilment, and understanding and knowledge. These basic affections are not contingent desires. Now if we have a capacity by which we move ourselves, and we certainly have it, then these inclinations and their objects are properly understood as constitutive of it. And they are properly appreciated as the primary necessary ends of our life. I believe that I am not making a novel argument here - one finds it in Aristotle, Aquinas and many others -, but this does seem to bear repeating. Now these natural affections or primary intrinsic goods are the indisputable source of most of our intentions, choices or decisions.

Second, my view that the will is a distinctive capacity of the mind implies that an intention is a distinctive mental state. And this view of intention is distinctive in the literature. Philosophical works on intention exhibit, roughly, two broad conceptual orientations. On the one hand, there is the tendency to treat intention as a species of one of the garden-variety mental states or a combination thereof. Views of intention as a higher-order desire (Frankfurt, 1988), as a special kind of belief (Hampshire & Hart, 1958; Harman, 1976; Velleman, 1999; Langton, 2004), as an all-out or unconditional desirability judgement (Davidson, 2001[1978]; Tenenbaum, 2007), as a judgement that is distinctive in that it categorically affirms a personal action (Ricoeur, 2009), or as a desire-belief pair (Davidson, 2001[1963]; Audi, 1993) fall into this theoretical category. On the other hand, there is the tendency to conceive of intention as a *sui generis* mental attitude. Representative views of this kind include Bratman's (1987) account of intention as a distinctive pro-attitude with a characteristic volitional and reasoning-centred commitment, and Raz's (2022) account of intention as a distinctive kind of mental state, not consisting in a combination of other states, but having an evaluative belief as a constitutive component.

Now my view is that an intention is not reducible to a belief or a desire (or a pair thereof) but that it sublates (i.e. presupposes and transcends) them. It obviously belongs to the second category of conceptual tendencies. But it is distinctive in showing how intention properly relates to the garden-variety mental states, such as belief or judgment and desire. This view, which I develop in Chapter 6, rejects the treatment of intention as a species of one of these states or a combination thereof. This treatment is mistaken because, as it has been argued in one way or

another by some philosophers (e.g. Pfänder, 1967, pp. 3-21; Carey, 2017, pp. 72-73), you can believe or judge that you have most reason to do something and yet not intend to do it or intend to do something else. Similarly, you can consciously have a predominant desire to do something without thereby intending to do it. Moreover, it is practically irrational (and presumably criticisable) to entertain two contradictory intentions, but not to have two conflicting desires. Finally, there are cases where we cannot form judgements or beliefs about what we have (most) reason to do but we can (and normally do) form intentions to act. This is what happens, for instance, in the so-called Buridan's ass situations, in which the reasons bearing on the agent's normative judgements are equally weighed or incommensurable. The fact that we are typically able to decide (and hence form intentions about) what to do even when the courses of action open to us strike us as being on a par reasons-wise suggests that an intention is distinct from a judgement or a belief. I develop (some of) these points in the next chapter.

Third, my view of intention is also distinctive for its treatment of the kind of commitment involved in the state of intention. Broadly speaking, having an intention to do something is, contrary to having a desire, being settled or decided upon doing it. Most philosophers working on intention have in common this conception of intention as a special kind of commitment and the belief that a full characterisation of this commitment is one of the main jobs of a theory of intention (e.g. Harman, 1976; Bratman, 1987; Velleman, 1999; McCann, 1991; Mele, 1992a; Frankfurt, 2006; Korsgaard, 2009). However, they disagree on the nature of the commitment characteristic of an intention. In the previous three chapters, I argued against Bratman's view that this commitment involves, *inter alia*, a reasoning-centred disposition (that is, a disposition to reason from one's retained intention to yet further intentions), against Velleman's view that an intention is a cognitive commitment to the truth of what it represents, and against Korsgaard's view that willing is a universalisability commitment that is essentially moral.

Let me spell out one reason for my opposition to Bratman's and Velleman's construals of the commitment involved in an intention. Bratman characterises this commitment partly in terms of intentions being conduct-controllers in the interim between their acquisition and their execution in action. This is part of what he calls the volitional dimension of the commitment characteristic of an intention. And Velleman holds that intentions are cognitive commitments to the truth of the act-descriptions that they represent. Now there are facts about our experiences of intentions that

do not seem to cohere with these notions of commitment. We often forget our intentions; we often change our mind about our intentions; and (without any change of mind) we often fail to execute them; and when we execute them, we often fail to carry it through. These are cases of the failure of intentions to control behaviours and lead to their execution, or to embody the truth of the act-descriptions that they represent. If intentions characteristically involve a special kind of commitment, this must be the case even for unfulfilled (e.g. forgotten, flouted, rejected, etc.) intentions. We need a characterisation of the commitment involved in an intention that makes sense of this fact. Bratman's and Velleman's conceptions of it appear to fail on this score.

My proposal is this: to grasp the nature of an intention and its characteristic commitment we need to look inside the structure of our consciousness and, specifically, in the distinctive, enquiry-guided pattern of the four levels of conscious intentionality: experience, understanding, judgement and decision. Two lessons can be drawn from this pattern. The first, which is about the nature and structure of an intention, is that in order for one to decide or intend, one must judge; in order for one to judge, one must understand; and in order for one to understand, one must experience. Deciding or intending thus presupposes and goes beyond judging, understanding and experiencing. As such, it is not reducible to any or all of them. This is why I said earlier that, on my view, an intention is not reducible to a belief or a desire (or a pair thereof) but that it sublates (i.e. presupposes and transcends) them. The second lesson is about the commitment involved in an intention as such a complex mental entity. On each of the four levels in the pattern there are distinctive questions that arise and that need to be settled. This is not surprising because as, I noted in Section 1, the structure of our consciousness is inquisitive. Now specific to the fourth or volitional level are questions about whether to affect ourselves and the world, whether to change ourselves and the world, in a desired or intended manner. These are fundamentally existential questions. To settle them, and hence form an intention, is then to make an existential commitment.

But what exactly is involved in such a commitment? To spell this out, I will help myself to the part of Korsgaard's conception of willing on which I am in agreement with her. On Korsgaard's view, willing or intending to do something is committing oneself to do it in all the circumstances that are relevantly similar to the expected ones and unless there is a good reason why not (2006, p. 61). Suppose, for example, that I now will or intend to work on my thesis tomorrow from

10am to 8pm. In so willing, I anticipate the circumstances that must obtain at 10am in order for me to start working on my thesis. And I commit myself to do so in the circumstances relevantly similar to those I expect and provided that there is good reason not to. If, for instance, I wake up with a debilitating migraine attack, this will be a good reason for me to give up my project. And giving it up for this reason will not flout the universal commitment characteristic of my willing because it is consistent with the condition attached to this commitment. But if I give up my intention because, say, I am feeling lazy, then I have violated my volitional commitment. And if I am prepared to give up my project for any consideration that strikes me as a reason, then I have not really willed, or committed myself to, anything (p. 62). I share this much with Korsgaard. I do not agree with her description of the commitment involved in willing in terms of the will being subject to ‘a formal version of the categorical imperative, the principle that our maxims must be willed as universal laws’ (p. 61). To me, willing involves this sort of ‘universal commitment’ because that is just what willing is intrinsically about, as opposed to, say, desiring, hoping or wishing.

We may further characterise this commitment in terms of self-determination and self-binding. In willing or forming an intention, the agent perceives a course of action or a state of affairs, and grants it, so to speak, his authoritative approval, or imposes on it the stamp of to-be-realised, thereby practically resolving to bring it about. The agent attributes to himself the direct or indirect responsibility for the realisation of the projected state of affairs. He makes an existential commitment to change the world, to bring into being something which, without him, will not exist. But he also makes a commitment to change himself, to make himself what he wants to be.

Now how does this characterisation of the commitment involved in an intention in terms of a distinctive existential commitment make sense of unfulfilled (e.g. forgotten, flouted, rejected, etc.) intentions? On my view, the stability of intentions and their power to lead to the intended actions are a function of the agent’s individual’s character, habits and personality, on the one hand, and of the strength of the reasons or needs for which the agent forms the intentions. I argued for this, with some evidence from empirical social psychology, in Chapter Two (subsection 2.2). An agent’s psychological and other characteristics may interfere with his intentions in different ways, and this is consistent with those intentions embodying an existential

commitment when they were formed. So cases of the failure of intentions to lead to the intended actions do not pose a threat to my view.

4

I began this chapter with a couple of related claims. The first is that what we do crucially depends on what we think and what we want, or on our reason and our will. And the second is that understanding human action requires grasping the sort of creature that a human being is, on the one hand, and the nature and relation of reason and will, on the other. This chapter addressed only the issues of the nature of the human being (mainly by analysing the structure of human consciousness) and of practical reason and the will. I argued that (i) our consciousness is structured in such a way that we are characteristically inquisitive about, and hence active with respect to, our perceptual and mental experiences; that (ii) practical reason is our capacity for determining, through evaluative reasoning, what we have most reason to do, and that this reasoning concludes in a value judgement about action; and that (iii) the will is a distinctive capacity of the mind, and that the commitment involved in a state of intention is an existential commitment.

This chapter also flagged up two issues for treatment in the next chapter. One is a defence of the claim that practical reasoning concludes in a judgement of what an agent has most reason to do, and the other is a discussion of the claim that intention is a distinctive mental state that sublates belief and desire. These two issues are part of the larger problem of the relation between reason and the will in the constitution of human action. And this problem is the object of the last chapter.

Chapter Six: Action, Reason and Will: Intention Sublates Judgement and Desire

The German “aufheben” (“to sublate” in English) has a twofold meaning in the language: it equally means “to keep,” “to ‘preserve’,” and “to cause to cease,” “to put an end to”. Even “to preserve” already includes a negative note, namely that something, in order to be retained, is removed from its immediacy and hence from an existence which is open to external influences. - That which is sublated is thus something at the same time preserved, something that has lost its immediacy but has not come to nothing for that (Hegel, 2010, pp. 81-82).

This chapter is a continuation and extension of the previous. The previous chapter dealt with the nature of practical reason and the will as distinct mental capacities that are most central to human agency. I argued that practical reason is the capacity to settle, through evaluative reasoning, the question of what we have most reason to do, and that the output of practical reasoning is an evaluative belief or judgement about what we ought to do. I also argued that the will is our affective capacity for intentions, and that an intention is a distinctive mental state that involves an existential commitment to a course of action. A defence of this view of practical reasoning and an explanation of the distinctiveness of the state of intention and its connection to practical judgement and desire were marked out as concerns for this chapter.

In this chapter, I take up the problem of the relation of reason and will in the constitution of human action. A range of apparently puzzling phenomena about our agency arise at the interface of reason and will, so that any account of the nature and relation of reason and will must accommodate them in order to be acceptable. I first briefly survey these phenomena in the first section, and then go on to address them in the subsequent two sections. All in all, I maintain the dualist account of practical reason and the will developed in the previous chapter, and argue that the relation between them is circular and dynamic and that it is best characterised as sublation (dependence and transcendence by affirmation or negation). I show how this account capably accommodates the said puzzling phenomena, and how they pose threats to views of the will as practical reason.

Let me begin with an overview of those phenomena. On the one hand, there are phenomena that seem to favour a separation between practical reason and the will. Two such phenomena are worth noting here, at least because they have a very long history in philosophical debates of agency, freedom and morality. These are: first, our ability to decide and act in Buridan's Ass cases, where the reasons pertaining to the practical judgement about alternative options are equally balanced, inconclusive, ambiguous, or incommensurable, so that they cannot warrant any judgement of which option is more choice-worthy. Second, cases of akrasia or weakness of the will, where we act against our best judgement, seem to attest a separation between reason and will. In such cases, though we are able to form the judgement that, all things considered, we have most reason to do one thing, we end up doing something different. Now based, as it is, on a dualist view of reason and will and a narrow view of practical reasoning, my account of the constitution of action looks immediately able to handle these phenomena.

On the other hand, other phenomena seem to push in the opposite direction. Two of them are also worth mentioning here. First, there is the phenomenon of instrumental rationality that agents typically display in their akratic actions or in the pursuit of goals that they do not judge to be all-things-considered worthwhile. As Wallace (2006) remarks, individuals 'sometimes exhibit great intelligence and skill in executing plans that they view as dubious or questionable,' - a kind of intelligence that can plausibly be regarded 'as a form of instrumental rationality' (pp. 82-3) at the service of their chosen ends. For illustrative purposes, consider the practical intelligence exhibited by a married man who intends, against his better judgement, to have an extramarital sexual intercourse and takes all the necessary precautions to ensure that his family and friends do not find it out.

Now Wallace brought up the phenomenon of instrumental rationality in akratic actions in the specific context of debates about the normativity of the instrumental principle as a principle of practical reason. But this phenomenon is also relevant to discussions of the relation of practical reason and will. In fact, Hieronymi (2009) takes the phenomenon of rationality in akratic actions as a serious difficulty facing any account that separates willing from reasoning. This account, she claims, 'will need the ability to employ reason for the purpose of determining exactly how to

successfully execute our sometimes-complex akratic actions’ (p. 216). In a similar vein, Watson (2007) has also argued that a proper treatment of the phenomenon of akrasia seems to caution against a (radical) disconnection of the will from reason. He warns that, by operating such a disconnection, we might face trouble explaining what is, after all, weak in akratic behaviour (p. 180).

Second, there are the so-called facts about our answerability for our willings. Hieronymi (2009) provides these three facts: first, the activities, roles and failures of the will are those of the person herself or as a whole (p. 203). Second, the person is answerable for the activities of her will in a quite particular way: she can rightly be asked for her reasons for those activities (p. 203). Third, an agent’s decisions, intentions, and intentional actions are adequate grounds for normative or evaluative responses - e.g. love and hate, liking and disliking, gratitude and resentment - towards him as a person because they reveal his character or personality traits in distinctive ways. (p. 204). Hieronymi claims that it is hard to understand this answerability if the will is not understood as practical reason (p. 203).

Finally, there are concerns that seem to apply to any accounts that explain human agency and autonomy solely in terms of reason and will. One such concern questions the legitimacy of the duopoly of reason and will in the business of freedom explanations. For example, Murdoch accused Hampshire’s book *Freedom of the Individual* of ‘working with a strict dualism of will and reason’ in its explanation of human freedom, and hence of ‘being unable to properly accommodate the imagination’ (Stern, 2022, p. 13). This criticism can just as well be applied to accounts of human agency and autonomy, such as all those discussed in this thesis, but with the broad range of human capacities in mind. (One thing worth noting in this connection is that it is a common criticism of the analytic philosophy of action that it completely neglects perception and its role in action.) Another concern, which is related to the first, is the complaint against attempts to explain human agency through the lens of a faculty psychology. Lonergan (2004), for instance, opposes such attempts on the ground that this psychology ‘divides man up’ into reason, will, emotion, etc., and generates ‘unresolved problems of priority and rank’ (p. 37), as famously illustrated by the medieval disputations between intellectualists and voluntarists or modern-day debates between cognitivists and non-cognitivists about moral judgements and motivation. Let us call these two concerns the problems of the strict dualism of reason and will.

Addressing these last two sets of problems - namely, facts about our answerability for our willings and the phenomenon of rationality in akratic actions, on the one hand, and the problems of the strict dualism of reason and will, on the other - will afford me an opportunity to develop further my accounts of the nature and relation of reason and will. I believe that successfully addressing these issues requires treating reason and will as distinct, though related, capacities of our nature. And this is what I am going to do in the rest of this chapter, starting with facts about our answerability for our willings and the phenomenon of rationality in akratic actions (Section 2), and ending with the problems of the strict dualism of reason and will (Section 3).

2

Let me begin with the facts about our answerability for our willings and the phenomenon of rationality in akratic actions. Regarding our volitional answerability, I first motivate a dualist account of reason and will by arguing that some of our ordinary ways of willing or forming intentions do not involve practical reasoning and that, for this reason, they pose serious threats to monist views like Hieronymi's. And then I develop a species of the dualist account of reason and will, on which the will drives and sublates (i.e. presupposes and transcends) practical reason, and show how it adequately makes sense of our volitional answerability (2.1). Moving on to the phenomenon of rationality in akrasia, I briefly argue that it is a problem only for dualist views of reason and will as completely unrelated. I then bring up a variety of akrasia that consists in the revision of intention in the face of temptation, and argue that it is hard to explain on an account of will as practical reason (2.2). All in all, my arguments in this subsection support the traditional philosophical view that we need reason and the will as distinct capacities to account for our ability to choose, stick by our intentions, and act freely and spontaneously 'in a complex world that is full of uncertainties and temptations' (Holton, 2009, p. x).

2.1. Hieronymi (2009) claims that it is hard to understand our volitional answerability if the will is not understood as practical reason (p. 203). She interprets the Kantian claim that the will is practical reason or 'reason in its practical employment' along the following lines: willing is an exercise of our capacity for reasoning - our capacity to raise, deliberate about, and answer questions. On this interpretation, 'willing [i.e. making a decision or forming an intention] to ϕ is

or centrally involves settling the question of whether to ϕ , where that settling is the kind of activity done for reasons' (p. 204). I think, however, that our best hope for making sense of our answerability for our willings is a dualist account of reason and will, and not a monist one like Hieronymi's.

To see why, let us consider some of the problems facing Hieronymi's variant of the Kantian monism about practical reason and will. Hieronymi divides our capacity for reason by distinguishing between practical reason and theoretical reason. She further divides theoretical reason into what we may call theoretically oriented theoretical reason (or theoretical reason for theoretical subject-matters) and practically oriented theoretical reason (or theoretical reason for practical subject-matters). Regarding action, it appears that only practical reason and theoretical reason for practical subject-matters have a direct bearing on its determination. Now Hieronymi construes practical reasoning as reasoning about whether to act - it is directed at the question of whether to do something. It concludes directly in an intention to act. But she leaves it unclear what the practical syllogism is or what the premises are of a reasoning that will conclude with an answer to the question of whether to A or not, where A is an action description.

As to theoretical reasoning about practical subject-matters, Hieronymi seems to give it two different formulations. The first formulation of this reasoning is that it is a reasoning about what you have most reason to do or what you ought to do: 'Judging that you have most reason to ϕ [...] is a piece of theoretical reasoning, about the reasons for acting one has (p, 207). And the second formulation is that this is a reasoning about your reasons for action: this reasoning 'results in a belief or judgements about one's reasons for action' (p. 207). These two formulations seem different: you reach a judgement about what you ought to do by reasoning about the reasons that bear on that judgement. But your reasoning about your reasons need not issue directly in a judgment about what you ought to do. It may consist mainly in isolating the reasons that should bear on your judgement about what you ought to do. As Hieronymi notes, '[w]e certainly expect a person to be guided, in answering the question of whether to act, by the reasons she acknowledges, when reasoning about her reasons for action' (p. 207). Similarly, we may add here, in determining what one ought to do, one is expected to be guided by the reasons that he has isolated as worthy of bearing on that judgement. So reasoning about what you have

most reason to do (or forming a judgement about what you ought to do) and reasoning about your reasons (or forming a judgement about your reasons) are not the same thing.

But perhaps a more troubling issue for the monist view that willing is or centrally involves settling a practical question concerns its implication that intending requires an episode of practical reasoning. Now it does not seem to be the case that each of our intentions results from such an episode. Suppose that I have (formed) an intention. What does this involve? How do we come to have intentions? There are perhaps two different ways we come to have intentions: passively (without doing anything) and actively (by a mental action - forming an intention, choosing or deciding).

Passive intention formation or acquisition happens without us making any explicit decision. In such cases, the agent just finds himself committed to the corresponding course of action without consciously deciding for it or against its alternatives. This is the case for many actions done out of habit. For example, habitually brushing my teeth in the morning is an intentional action that is not preceded by a prior explicit decision to do so. Apart from habituation as a passive mode of intention formation, there is also post-hypnotic suggestion, which is an extreme case of passive intention acquisition. For example, through a suggestion like 'When you wake up, you will go to the kitchen and make yourself a cup of tea and a sandwich' one may acquire the corresponding intention. This mode of intention acquisition raises serious moral and legal issues (especially in cases where the resulting behaviours are of a criminal nature). When an individual acts in accordance with an instruction given - or an 'intention' formed - in hypnosis, does he act intentionally? Is he morally or legally accountable for such an action? Intention formation through habituation does not raise such questions presumably because the state of consciousness underlying such a mode is not one in which the individual is a passive or helpless bystander. These differences notwithstanding, it appears that the passive mode of intention formation or acquisition as a whole does not, in essence, involve (practical) reasoning. And so this mode poses a threat to Hieronymi's account.

Now Hieronymi may quite plausibly be interpreted as talking about instances of active intention formation only. Active intention formation happens typically through practical reasoning - in the form of conscious choices or decisions. And it is the paradigmatic case of intention formation. It

is the paradigmatic mode because we are self-conscious and rational creatures: we are the sort of creatures who typically think about their experiences and are capable of higher-order responses to those experiences. And so when we become aware of our passively acquired intentions, we often wonder whether they make sense, thereby making it possible for us to rationally retain or discard them. It is because we are normally this reflective and active with respect to our intentions and other motivational states that we are typically held responsible for them.

But that we paradigmatically form intentions through practical reasoning does not in itself support or entail the view that the will is practical reason or the view that practical reasoning concludes in intentions. If anything, it simply raises again the question of the proper boundaries of practical reason(ing). On a narrow conception that I am defending here,

all reasoning is concerned with the formation or modification of beliefs, and practical reasoning in particular is concerned with the formation or modification of normative beliefs about action. The adjustments in our intentions that result from such reflection are not themselves conclusions of reasoning [...]. Agents who have resolved the question of what they ought or have reason to do still have a question to settle, about what they are going to do (Wallace & Kieseewetter, 2024).

On my account, answering this further question - of what one is going to do - is not the business of reasoning: practical reasoning has run its course once it has yielded a judgement about what ought to be done. Settling the question of whether to do it, and hence forming an intention, is the sole business of the will.

A good reason to believe that intention formation is the will's own work and, accordingly, to reject the conception of will as practical reason, is that not every active instance of willing, deciding or intending occurs through an episode of practical reasoning or deliberation. Hieronymi claims that everyday willing does not require practically oriented theoretical reasoning - that is, the formation of a judgement about your reasons or about what you ought to do -, but it requires practical reasoning or reasoning whether to act (p. 207). But this strikes me as implausible. We often form intentions to act not only without forming judgements about reasons but also without reasoning whether to act. As I am walking through the aisles of the supermarket, I hear a certain song start playing, and upon hearing it I form an intention to play it

when I get home. I did not form any belief or judgement about my reasons for playing the song at home, nor did I raise and settle the question whether to play the song at home. Presumably, I formed my intention merely on the basis of the desire, feeling or emotion aroused in me by my hearing the song play. Willings or intentions such as this are spontaneous.

Let us consider another example. 'Tea or coffee?', you ask me as you lead me into your living room. 'Tea,' I instantly reply, thereby letting you know my intention. I have formed this intention actively, or I have settled the question of whether to take tea or coffee, but without engaging in practical reasoning. And it need not be the case that this intention is a longstanding commitment formed during an episode of deliberation. We are the sort of creatures that are capable of such pure spontaneity. Now in light of these facts about how we form intentions, we need to distinguish two ways or levels of settling a practical question in our lives: a rational settling and a volitional settling. A rational settling issues in a judgement about what to do. A volitional settling issues in an attitudinal commitment to a course of action. The two are importantly distinct forms of question-settling. If so, then the view that the will is practical reason is false.

Hieronymi contends, however, that the fact of spontaneous intentions 'does not itself tell against the view of the will as practical reason. After all, we also spontaneously form judgments and come to believe things without prior conscious thought' (p. 216). The point here seems to be that reason is the capacity for judging and believing, and yet we often form judgements or acquire beliefs without reasoning. But this fact does not discredit the conception of reason as capacity for judgement and belief. Now, similarly, there is nothing awkward about holding that the will is the capacity for practical reasoning while recognising that we often decide or intend to do things without reasoning.

I think that the point needs to be made differently. I believe that a more plausible way to put it is to say that only a creature with cognitive abilities can have beliefs and judgements. Or that the fact of a creature having beliefs and judgements presupposes its possession of cognitive abilities. An identity relation does not follow from this: for example, in order for me to have perceptual beliefs or memorial beliefs, I must have cognitive abilities like the capacity for reason. But it does not follow that the capacity for reason is the capacity for perceptual beliefs or memorial

beliefs. Perceptual beliefs come for perceptual experiences and the exercise of the capacity for perception, and not the capacity for reason. Similarly, we may hold that the fact of a creature having willings presupposes its possession of the capacity for reasoning, although willing and reasoning could be different mental activities. And this presupposition should be enough to ground facts about our answerability for our willings.

Presupposition or dependence is, however, just one constitutive element of the circular and dynamic relation that exists between our faculties of reason and will. This relation, with its other elements, is most plausibly characterised as *sublation* in the relevant sense. Now there are two concepts of sublation in the philosophical literature, both of which are inadequate for the relation between reason and will. The first concept of sublation originated in the philosophy of Hegel, where it refers to the dialectical process in which conflicting positions or forces are preserved and brought into a higher-level integration or unification. As Hegel puts it in *The Science of Logic*, '[s]omething is sublated only in so far as it has entered into unity with its opposite' (2010, p. 82). The Hegelian concept of sublation is inadequate because there is just no sense in which the faculties of reason and the will can justifiably be viewed as being in dialectical conflict or opposition. (As I will argue shortly, the will - specifically, the natural desire for knowledge and truth - is what drives the activities of reason.) This should not, however, be taken as a criticism of Hegel, because he presumably did not develop his concept of sublation with the relation of reason and the will in mind or to address the question of the relation between practical reasoning and the formation of intention.

The second concept of sublation is owed to Karl Rahner's interpretation of Hegel and gets a more extensive treatment in Bernard Lonergan's philosophical and theological thought. Unlike Hegel, Lonergan uses the term 'sublation' to account for the relation between the conscious and intentional operations of experiencing, understanding, judging and deciding sketched in Section 1 of Chapter 5. He uses the phrase 'sublation of A by B' in the sense of B presupposing, preserving, complementing and going beyond A (2016, pp. 72-3). In this sense, to say that the pattern of these operations (sorted in ascending order) is sublating is to say that each higher operation requires and retains (all the properties of) the lower one(s) but goes beyond it (them) by introducing something entirely new and distinct and by making this added element a new basis of operation (2017, p. 241). In this pattern, experiencing is sublated (that is, presupposed,

complemented and transcended) by understanding. Experiencing and understanding are both sublated by judging. And experiencing, understanding and judging are all sublated by deciding.

To put flesh on the bones of Lonergan's concept of sublation, I will consider an example. Suppose that, as I am working on this thesis in my room, I hear loud noises coming from the next room. I stop typing for a bit and try and listen intently, wondering what the noises could be about. After a few seconds, I have the realisation that there is a fight going on in that room. Listening yet more intently and analysing the perceived voices, I form the judgement that it is indeed a fight between a young man and a young woman, and actually a bad fight. At this point, a question arises in my mind whether to ignore the fight and go back to my work or instead do something - say, go and knock on the door. I decide to go and knock, and instantly start for the door.

In this scenario, I have had a perceptual experience of loud noises, interpreted or understood them to signal a fight, judged that it was a bad fight between two people, and then decided to go and knock on their door. These four conscious events are closely related in a particular way: they are interdependent: my understanding that there are loud noises in my neighbour's room because of a fight sublates my perception of those noises. That is, my understanding of the nature of the noises presupposes a sense experience of them and preserves the perceptual data of this experience, but it also adds meaning or intelligibility to them. And in adding this new element, my understanding complements and goes beyond my sense experience. And the meaning or intelligibility grasped by the operation of understanding is the basis of my judgement that there is indeed a bad fight going on in the next room. My judgement, in turn, goes beyond my understanding by affixing the seal of reflective affirmation to the intelligibility of my perceptual experience grasped by that operation. And, finally, my deliberative decision goes beyond my judgement by using it as a ground for raising and settling a question about action.

Lonergan's concept of sublation thus posits presupposition, preservation, complementation and transcendence as the constitutive elements of the relation between the operations of experience, understanding, judgement and decision. I think, however, that this concept is inapt for the relation between (some of) these operations in a range of cases. One such case is the case of non-veridical perception, as in Macbeth's hallucination of a dagger. In hallucinating a dagger,

Macbeth, it may be said, at least ‘perceptually believes that there is a dagger before him’ (Audi, 1993, p. 93). Now suppose that Macbeth ended up dismissing this belief, judging correctly that it originated from a mere hallucination. In this case, Macbeth’s judgement would presuppose and transcend his perceptual belief but without preserving, let alone complementing, it in any way whatsoever. Here, the judgement transcends the belief by negating it. If so, it appears that preservation and complementation are not necessary for the sublation of one operation by another.

Another case that poses a problem for Lonergan’s concept of sublation comes from the practical sphere: the case of *akrasia*. In one form of *akrasia*, the agent’s practical judgement (or reason) and his decision (or will) go apart. For example, I judge that, instead of going clubbing with my friends tonight, I should stay home in order to finish this chapter, but against this I decide to go and have some fun. This decision presupposes my judgement that I should work on my thesis and transcends it in a way that neither preserves nor complements the content of that judgement. It is a clear instance of transcendence by negation. So this practical case, like the theoretical case above, suggests that the relation of sublation between conscious and intentional operations of the mind doesn’t require the preservation and complementation of a lower operation by a higher one.

In light of the foregoing, I propose here a concept of sublation as a relation of presupposition or dependence and transcendence (by affirmation or negation) between two or more entities. It is this concept that, I think, best captures the circular and dynamic relation that exists between the reason and will of human subjects. But this picture of sublation needs a few more brushes. We need to add that, in addition to the will presupposing and transcending reason, the will also drives reason¹. There are at least two ways in which our will - that is, what we want, love or care about - motivates the activities of our reason. First, the will drives reason as the motivator of rational enquiry. Our reasoning is driven by our natural desire to know, and this desire is an inclination of the will as an appetitive capacity of our nature. Without this desire it would not be possible for us to commit ourselves to scientific, philosophical and practical pursuits, or to seek

¹ This talk of reason and will as doing things is only metaphorical. I do not take reason and will as homunculi but simply as ‘factors in the acting of the single, unitary agent, the acting person’ (Finnis, 1998, p. 65).

to understand, judge and decide. But as the motivator of rational enquiry the will can - and often does - also have a corrupting effect upon reason and its judgements. This is typically the case when our preference for, or aversion to, certain states of affairs inclines us to select and/or interpret the data on which we base our judgements in ways conducive to desirable conclusions. Second, in addition to motivating reasoning - both theoretical and practical -, the will also furnishes practical reasoning with its inputs: our practical reasoning not only is driven by our deep-seated urge to know what to do, but also normally operates within the basic structure of desires, preferences and cares that we happen to have².

There is, then, a complex relation between reason and will in human psychology. In sum, I have argued for a dualist account of reason and will, on which the will drives reason, on the one hand, and the will sublates reason (i.e. presupposes reason and transcends it by affirmation or negation), on the other. This account explains why intention formation or acquisition need not involve practical reasoning, and why it is the business of the will alone though it presupposes the capacity for practical reason. Now thanks to the notion of sublation as dependence (or presupposition) and transcendence, this account also makes sense of why we are distinctively answerable for our willings - or the facts that your willings are yours as a person, that you can be rightly asked for your reasons for your willings, and that your willings open you to a certain characteristic range of evaluations (Hieronymi, 2009, p. 204). For Hieronymi, to explain volitional answerability, which is the accountability of the agent as a whole, willing must be an activity of the whole person (rather than of some sub-personal mechanism). And it can be such an activity only if it is an exercise of reason, because exercises of reason are activities of the whole person (p. 205). If we grant, for the sake of argument, that only reasoning is an activity of the whole person, it will not follow that willing must be reasoning in order to count as such an activity. For it would suffice for willing to sublate reasoning in order to satisfy the condition for attributability to the whole person. So, contrary to Hieronymi's claim, it is not hard to grasp volitional answerability if the will is not construed as practical reason.

2.2. Neither is it hard for a dualist account of reason and will to accommodate the phenomenon of rationality in akratic actions. Recall that this is the phenomenon whereby reason is put at the

² For an extended discussion of these issues, see Steel (2010), 'The Effect of the Will on Judgement: Thomas Aquinas on Faith and Prudence,' pp. 78-86.

service of the pursuit of rationally unendorsed ends. Now accounting for this phenomenon would pose a problem only to those dualist views that construe the relation of reason and will in the same terms as Rudyard Kipling's poem 'The Ballad of East and West' understands the relation between East and West: 'East is East, and West is West, and never the twain shall meet'. But this is clearly the extreme opposite of the sublation-based dualism that I am defending here. On my account it is not at all mysterious how an akratic agent is capable of instrumental rationality: we often settle on certain goals against our best judgement and then reflectively figure out how to achieve them, what obstacles (if any) to anticipate and overcome, etc. In these cases, practical reason is demoted from playing both substantive and instrumental roles to playing only an instrumental role to desired ends that it disapproves of but that entice the will. This demotion is possible because, as I argued above, the will drives and sublates practical reason.

Now putting aside the phenomenon of rationality displayed in akratic actions, I think that akrasia itself is a veritable challenge to Hieronymi's own account of the will as practical reason. What I have in mind here is not the variety of akrasia that consists in intending and acting against one's best judgement, but the one that consists in the failure to act in accordance with one's intentions. This failure could be explained by a variety of factors, not all of which lead to akratic behaviours. And so not all of them are the phenomenon of interest to us here. Suppose I decide now to hit the gym tomorrow from 4pm to 5pm. Suppose, further, that tomorrow I end up not doing what I decided. What could explain this state of affairs? It could be that I changed my mind, having realised that I had a stronger reason to do something else, something more important, during that time. It could also be that something outside my control (say, a bad migraine attack) prevented me from carrying out my intention. Or that, in some mysterious way or for some mysterious reason, the plan to go to the gym skipped my mind during the day. Or that I took a nap after lunch and failed to wake up in time for the gym. It could also be that, having spent the day cooking, cleaning and doing my laundry, I just had no energy left for the gym. A final possibility is the phenomenon of giving into the temptation to do something that my intention rules out. Here, I merely succumbed to the temptation to stay in bed and watch a movie instead of going to the gym as planned. And in so doing, I revise or reverse, unreasonably, my earlier decision, presumably under the pressure of the desire or pleasure to do something else.

It is this phenomenon of reconsidering an intention in the face of temptation that relates to the variety of akrasia of interest to us here. It is of interest because how temptation works, but also how to resist it, teaches us something about the nature and relation of reason and will. However, recipes for resistance are outside the scope of this chapter. I propose two forms of succumbing to temptation: *instantaneous succumbing* and *progressive succumbing*. An instance of instantaneous succumbing to temptation is the one just sketched above: as the time of going to the gym approaches, I feel a stronger inclination to stay in bed and watch a movie, anticipating the pleasure I will derive from so doing, than my inclination to stick to the plan of hitting the gym. And I succumb to that inclination, as evidenced by my staying in bed and my beginning to watch a movie. Or think of Jesus's third temptation in the wilderness: 'the devil took him to a very high mountain and showed him all the kingdoms of the world and their splendour. "All this I will give you," he said, "if you will bow down and worship me."' (Mt 4: 8-9). If Jesus had given into this temptation, enticed by Satan's promises of power and glory, his would have been an instance of instantaneous, clear-eyed succumbing to temptation. In the case of instantaneous succumbing, the agent consciously faces a choice between the option consistent with his intention and a tempting alternative, and he consciously embraces the alternative.

The second form of succumbing to temptation and revising one's intention is not at all as straightforward. To illustrate it, I will need to change the examples. Suppose that I have decided to spend the whole morning of tomorrow - in fact, the hours from 9am to 12pm - studying calculus and doing my problem set on differentiation in the Western Bank Library at the University of Sheffield before attending my afternoon lectures. When I arrive in the library at 8.50am, it may happen that I get involved into activities like reading online the controversy over who between Newton and Leibniz invented calculus, taking fresh air breaks, walking to and back from the maths section of the library, etc. But, while I go through these activities, I am secure in the belief that I am making progress with my calculus lecture notes and problem set - until it is time for my afternoon lectures and it dawns on me that I did not get any planned work done. I call this an instance of progressive succumbing to temptation and reconsideration of one's intention. In this case, the (metaphorical) devil uses his wiles to get me to revise my intention and succumb to his temptation without me suspecting it: he makes this happen over an

extended period of time, while feeding me with the illusion that I am steadfast in my plans and actually carrying them out.

I owe this form of temptation to Schwenkler (2024), who claims that it has hitherto been neglected by philosophers. But I disagree with his characterisation of it. Schwenkler distinguishes between what he calls *temptation to indecision* and *temptation to violation*: in the former, one is tempted to revise one's intention and do something else instead; in the latter, one is tempted to violate one's intention, or to act contrary to it, but without revising it (p. 448). These two cases of temptation correspond, respectively, to the two forms of succumbing to temptation that I have distinguished and explained above. Now I disagree with Schwenkler's characterisation on two counts. First, there is no distinction between indecision (as characterised here) and violation. In the case of temptation to indecision, the agent succumbs to temptation by revising irrationally, or for no good reason, his intention. But an irrational revision of an intention just is its violation. There is therefore no contrast between indecision and violation. This objection does not, however, spell much trouble for the distinction drawn by Schwenkler. For he might grant it and then proceed to replace the contrastive pair 'indecision-violation' with a suitable one, while maintaining the distinction between a form of temptation that involves a revision of intention and another form that does not. This distinction is the object of my second objection.

Second, I contend that there is a revision of intention in both cases. In the first case, the revision is instantaneous, and the agent is consciously aware of it; in the second, the revision is incremental, and the agent is not consciously aware of it. Much as 'coming to have an intention [is] a change that may take place so slowly or unnoticed that the agent cannot say when it happens' (Davidson, 2001, p. 89) - so that when we realise that we have an intention, we may have had it for a long time before -, coming to revoke an intention may occur in a similar fashion. Here, the agent who is progressively succumbing to temptation might believe that he has not reversed his intention. But that is simply because he is in the devil's grip of self-deception.

Schwenkler argues, however, that we should reject this interpretive appeal to self-deception for two reasons. Firstly, there is no self-deception about the status of my decision - that is, about its

revision or reconsideration - because the pattern of behaviour I am displaying in the library could result from 'my simply not *wanting* to act as I really have decided to, or [from] my more strongly wanting to do something else' (p. 451). That is, my behaviours may spring just from desires at odd with my intention, but not from its revocation. And, secondly, I should not be treated as being self-deceived because I may exhibit behaviours that indicate that I intend to do what I decided to do. For instance, that I memorised the questions in my problem set before leaving for the library, that I turned off my phones upon arriving in there; that during my fresh air breaks I was thinking about my differentiation questions; and that from time to time I recoiled at the realisation that I hadn't solved any question yet and tried harder to focus on reading the notes, etc. are all indicative of my commitment to study calculus and do the differentiation problem set.

But the behaviours captured in these two reasons furnished by Schwenkler are compatible with the unconscious and progressive revision of one's intention and with self-deception about its status. I repeatedly act on occurrent desires that are obviously contrary to my intention to study calculus and do my problem set, all the while believing that I have that intention. What reason do I have for this belief? Presumably, some of my behaviours - namely, my thinking about the problem set during my fresh air breaks, my trying harder to do it every time I realise I have been distracted away from it, etc. - give me a reason to believe that I (still) hold on to that intention. But behaving this way does not entail being motivated by the relevant intention. Because these behaviours could also plausibly be motivated and explained by other considerations, including the guilt or regret caused by my unconscious revision of my intention. Moreover, when I look back on my pattern of behaviour later - say, as I am on my way for my afternoon lectures -, I may find it hard to believe that I had kept my intention alive in the course of doing what I was doing. That is why this form of inconstancy in the face of temptation is more plausibly described as progressive succumbing to temptation.

There are, then, two different ways of succumbing to temptation by revising one's intention: instantaneous succumbing and progressive succumbing. Now I claim that both pose threats to Hieronymi's view of the will as practical reason. Hieronymi mentions in a footnote what is presumably a case of instantaneous succumbing to temptation by revising one's intention - she does not register any awareness of the other case - in order to show briefly why it is not a

challenge to her monist account. Hieronymi claims that the weakness of the will that consists in (instantly) revising one's intention in the face of temptation results from 'the inconstancy of practical reason' or the 'inconstancy of reason in its practical employment' (p. 218). But she does not explain how this inconstancy comes about, or what it exactly consists in, let alone how practical reason itself can overcome the temptation and lead to action.

But a similar view can be traced to Aquinas. According to Aquinas, the origin of inconstancy, or the failure to stick by our intentions, 'is in the appetite, for a man does not withdraw from a previous good purpose, except on account of something being inordinately pleasing to him'. That is, having a desire at odd with one's intention is a necessary ingredient of the revocation of intention. It is the primary cause of inconstancy. But this inconstancy or failure can be complete only because of 'a defect of reason, which is deceived [by the passions] in rejecting what before it had rightly accepted,' and it is so deceived 'due to its own weakness in not standing to the good purpose it has conceived'³. On Aquinas's account, the weakness of reason, rather than the pull and push of the passions, is the chief cause of the agent's inconstancy in the face of temptation. And so is the steadfastness of reason the cause of constancy. For him, the difference between the continent and the incontinent, and between the persevering person and the non-persevering person, is the condition of their reason: both the continent and the incontinent experience strong sexual urges, but in the former reason stands firm against the urges whereas in the latter it is deceived by them. And both the persevering person and the non-persevering person suffer severe pains in their course of action, but the former's reason is firm whereas the latter's is not. That is why Aquinas holds that constancy and inconstancy in the face of temptation both pertain to reason (ST.II-II. Q53.A5.C). But the weakness of reason then spills over the will, thereby causing the agent to succumb to temptation and act akratically.

This Aquinian explanation of the revision of intention in the face of temptation as inconstancy of practical reason cannot, however, be appropriated by Hieronymi to salvage her account of the will as practical reason. The reason for this is that Aquinas treats reason and will as two distinct rational faculties of the soul, and explains the will's revision of its intentions as a result of

³ In *Willing, Wanting and Waiting*, Holton presents a quite similar account of the revision of intentions in the face of temptations. He argues that when we reconsider our decisions, we are led to judge the tempting alternative as being worthwhile. Temptation both changes our desires and corrupts our judgments (2009, p. x).

reason's failure to resist the passions. This explanation is possible because Aquinas holds a dualist account of practical reason and the will. So it remains a potential challenge for Hieronymi's Kantian account to explain the possibility of the two variants of the kind of akrasia that consists in the revision of intention in the face of temptation (and also how to overcome it). It seems plausible that an explanation of this sort requires a dualist account of reason and will, but also Aquinas's style of treatment. But I disagree with Aquinas's intellectualist argument that the weakness of reason is the chief cause of inconstancy and revision of intention. Insofar as the will drives and sublates reason, as I have argued above, our succumbing to temptation is more plausibly construed as the effect of the weakness of the will enticed by the passions. But I will not pursue this line of argument any further.

3

I will now use the account of the circular and dynamic relation between reason and the will - that is, the view that the will both drives and sublates reason - developed in the previous section to address the concerns from the strict dualism of reason and will. Recall that these problems are the Murdoch-inspired charge that explanations of human agency in terms of reason and will exclude or cannot accommodate other relevant human capacities, and Lonergan's complaint that faculty psychology 'divides man up' and gives rise to 'unresolved problems of priority and rank'. I will first take up Lonergan's complaint (3.1) and then move on to the Murdoch-inspired charge about the accommodation of other constitutive capacities of human agency (3.2).

3.1. Lonergan's criticism calls into question the approach to action explanation adopted in this chapter and the previous one (but also this thesis). I started the previous chapter with the claim that what we do chiefly depends on what we think and what we want, or on our reason and our will, and that grasping human action requires grasping the nature and relation of reason and will. This claim presupposes a conception of ourselves (or our minds) at least partly in terms of our constitutive capacities and their uses, - a conception akin to the so-called faculty psychology. Now Lonergan complains that faculty psychology, which conceives of the mind as composed of functionally different capacities (e.g. reason, will, emotions, etc.), violates 'the unity and continuity of human consciousness' and raises insoluble questions:

A faculty psychology divides man up: it distinguishes intellect and will, sense perception and imagination, emotion and conation, only to leave us with unresolved problems of priority and rank. Is sense to be preferred to intellect, or intellect to sense? Is intellect to be preferred to will, or will to intellect? Is one to be a sensist, an intellectualist, or a voluntarist? (Lonergan, 2004, p. 37).

Lonergan contrasts faculty psychology with his analysis of conscious intentionality, which concerns the conscious and intentional operations underlying human cognition and action (e.g. experiencing, understanding, judging and deciding). And he argues that a shift from the former to the latter has the benefit of restoring and capturing the unity of consciousness and dissolving the above issues. In response to this, I think that faculty psychology, or any thinking about the mind in terms of powers or capacities, and Lonergan's intentionality analysis are not, in principle, opposed. For Lonergan's intentionality analysis itself assumes the existence of mental powers such as the capacities for sense perception, reason and will. It is because of the absence of any natural tension between faculty psychology and intentionality analysis that, in accounting for how action depends on what we think and what we want, I have used both the notion of distinct mental capacities and Lonergan's analysis of consciousness. And this suggests that, contrary to Lonergan's belief, adopting his intentionality analysis does not dispense from questions about the nature of mental powers and their relationship⁴.

On the other hand, Lonergan's criticism of any thinking of the mind (or self) in terms of powers is valid only if this thinking involves a commitment to the false notion that mental processes are starkly separate and inherently antagonistic. Such a conception is surely vulnerable to the charge that it is not consistent with the interconnectedness of mental phenomena. For example, the arousal of the emotion of fear in adult human beings is caused by the perception and evaluative cognition of threatening objects or events (e.g. a crawling snake or a sudden shooting) and typically involves the activation of standing motives (e.g. the desire to retreat or flee for safety). And cognition is also normally intertwined with motivation, as captured by my argument that the will drives reason. As this argument suggests, my conception of the mind in terms of distinct but interrelated capacities doesn't involve the problematic theoretical commitment, and so it is not unsettled by Lonergan's concern about the unity and continuity of human consciousness.

⁴ For extended arguments along these lines, see Crittenden (2012), *Reason, Will and Emotion: Defending the Greek Tradition Against Triune Consciousness*, pp. 83-96.

3.2. Now my account's notion of sublation (as dependence and transcendence by affirmation or negation) and its corresponding commitment to the interconnectedness of mental phenomena also thwarts the Murdoch-inspired charge that explanations of human agency in terms of reason and will exclude or cannot accommodate other relevant human capacities. The theorised relation of sublation between willing, judging, understanding and experiencing means that my account of human action in terms of reason and will is not strictly dualist; it implies the interdependence between reason, will and other capacities. Interestingly, this relation also means that intention is a distinctive mental state, one that sublates, and so is not reducible to, judgement or belief, desire, or a combination thereof. This view of intention is quite unique in many ways in the philosophical literature.

As I showed in Chapter 5 (section 3), philosophical works on intention exhibit, roughly, two broad conceptual orientations. On the one hand, there is the tendency to treat intention as a species of one of the garden-variety mental states or a combination thereof. On the other hand, there is the tendency to conceive of intention as a *sui generis* mental attitude. My account of intention as sublating desire, belief or judgement obviously belongs to the second category. But it is distinctive in showing how intention properly relates to the garden-variety mental states, such as belief or judgment and desire. Intention sublates desire and belief or judgement. That is, willing or intending presupposes and transcends desiring and believing or judging. I have argued at length in the previous section that will (intention) sublates reason (judgement). So I will not go into it again here.

The key takeaway from this argument is that the will is not practical reason, that willing or forming an intention is not the same as forming a judgment about what is the case, about value, or about what ought to be done. Forming an intention typically involves such judgements, but intention does not consist in any of these judgements. Intending sublates (i.e. presupposes and transcends) judging. As such, the act of forming factual, evaluative or ought judgements merely sets the stage for the active formation of intentions. Insofar as an intention essentially involves a distinctive existential commitment to a course of action, as I argued in the previous chapter, it appears that it cannot be reducible to a special kind of belief or judgement. Now what does deciding or intending (thus construed) add to judging? As I pointed in the previous chapter

(Section 3), deciding settles existential questions, which judging does not. Deciding adds an existential fiat, consent or resolve to evaluative judgements or beliefs about what to do.

In the rest of this section, I will focus on arguing that intention is not reducible to desire, that intention instead sublates desire. Now, at first sight, the reduction of intention to desire seems natural because there appears to be a close connection between intention and desire. Both intention and desire are pro-attitudes. And, as Lombardo remarked, ‘when talking about intention, we usually talk more frequently about what we *want* to accomplish than what we *intend* to accomplish’ (2013, p. 27). For example, after James has told John that he *intends* to travel to Italy next summer, John might report it to Patrick as follows: ‘James *wants* to travels to Italy next summer’. But is an intention really a desire? If not, how does intention exactly relate to desire? There are, perhaps, two best-known accounts of this relation in the philosophical literature: the view that intention is a special kind of desire and the view that intention only implies desire as a pro-attitude (in a specified sense).

Let’s begin with the view of intention as desire. One species of this view is the strongest-desire theory of intention. According to this theory, to have an intention regarding X consists in desiring X the most. Now there are various familiar arguments against the claim that intentions are nothing but desires, albeit predominant ones. I will here present quickly five of them. First, we typically distinguish clearly between desires and intentions: intentions are formed, whereas desires are not. They are acquired. And so one may experience a predominant desire for something without thereby having, or forming, a corresponding intention. Second, one can also form an intention to perform an action in spite of experiencing a predominant desire against it or in spite of not experiencing a desire for that action which surpasses in strength the desire against it. For example, a student may intend to stay home and revise for his exams even though he has a much stronger desire to go to a friend’s party instead. Third, similarly, one can form an intention not perform an action for which he has a very powerful desire. In spite of the violence or strength of the desire for the action, the agent is able to intend not to perform it. Fourth, the strongest-desire theory of intention is inconsistent with our intuitive conception and experience of ourselves as self-determining agents, as agents who can independently decide what to do, whatever the intensity of the desires we may or not be subject to at the moment. Finally, the very notion of willpower, or rather our folk-psychological understanding of it, suggests that willing or

intending cannot be a matter of desiring something the most. We need willpower to preserve our intentions against temptations, which normally come to us in the guise of predominant appetites or aversions⁵.

Having disposed of the view of intention as desire, I will now turn to the view that intention implies desire. This view is pretty complex, and needs some unpacking. Unpacking it takes us straight into the murky waters of the relation between desire and motivation. It begins with Nagel's distinction between unmotivated desire and motivated desire introduced in *The Possibility of Altruism* (1970). On this distinction, unmotivated desires are those desires that 'simply assail us' or 'simply come to us' unbidden - 'like the appetites and in certain cases the emotions' -, whereas motivated desires are those that 'are *arrived* at by decision and after deliberation' (p. 29). That these desires arrive by decision and after deliberation, comments Schueler (1995), does not mean that one decides to have such desires or that they simply arrive after deliberation. It means rather that these desires arise in virtue of the decision or deliberation (pp. 16-7). As such, motivated desires are subject to rational or motivational explanation, unlike unmotivated desires, which are produced (or caused) by physiological or biological states. On the basis of this distinction, Nagel concluded that the claims that a desire underlies every intentional action and that all motivation implies desires are true only if desires are taken to include both motivated and unmotivated desires. He considered it an open question whether an unmotivated desire always lies behind a motivated one or whether there can be motivation without any motivating desire (pp. 29-31).

Now is Nagel's distinction plausible? And is it crucial for understanding what desires are and their relation to motivation, intention and intentional action? Schueler (1995, pp. 20-25) takes this distinction to be crucial for understanding desires, and thinks that its only problem is non-exhaustivity. That is, the modality of simply happening to us and the modality of happening by decision and after deliberation are not the only two ways in which we come to have desires. Consider, for instance, my desire to visit my sister next summer. This desire is not, or need not be, held in virtue of any decision or deliberation. And it need not assail me like hunger, thirst or sexual urges. There is another way in which my desire to visit my sister is significantly different from these paradigm unmotivated desires: 'I can and do have reasons for wanting to visit my

⁵ In one form or another, some of these familiar arguments can be found in Pfänder (1967) and Carey (2017).

sister [...] next summer in a way in which it makes no sense to say that I have reasons for being hungry or thirsty [or lusty] (Schueler, 1995, p. 23). The issue, in other words, is that Nagel's distinction cannot account for the sense in which my desire to see my sister is (or can be) an unmotivated desire.

To address this issue, Schueler suggests that the distinction between unmotivated desires and motivated desires be redrawn in terms of, on the one hand, desires that resemble paradigm unmotivated desires, and, on the other, both desires that are merely creatures of some intention or decision (and associated beliefs) and desires that are not generated by such volitional states but are nevertheless held for a reason. That is, we should redefine unmotivated desires simply as ones not arrived at by decision and after deliberation or simply as those that are not rationally motivated, and re-construe motivated desires along the lines of the view that 'a desire is motivated if and only if the agent wants whatever it is that he wants for a reason' (p. 24). This re-construal allows for the possibility of motivated desires that are not generated by decisions or deliberations. So the class of motivated desires comprises both intention-generated motivated desires and those that are not so generated. According to Schueler, this way of redrawing the distinction between unmotivated and motivated desires enables us to see the sense in which my desire to see my sister can be an unmotivated desire even though there are reasons (unbeknownst to me) for me to want to visit her: if it assails me like a craving for chocolate, it is true that I want to visit my sister but it is not true that I want to visit her for a reason (pp. 24-25).

Having spelled out the distinction between motivated desires and unmotivated desires, we can now understand the view that intention implies desire (in some sense). If intention implies desire, and if we admit Schueler's distinction between motivated and unmotivated desires, it would seem that the relation between intention and desire will not hold for unmotivated desires but only for motivated desires, and exclusively those that are generated by intentions. Intention-generated desires, like all motivated desires, are not substantive desires; they do not refer to any particular psychological state. They are desires only in a placeholder sense, in which, it is claimed, 'desire indicates a motivational direction and nothing more' (Schueler, p. 34). And so the view that intention implies desire in this sense is not making a substantial claim about the relation between intention and desire. It is only making a claim about the intrinsically or independently motivating character of intentions.

Regarding the view of intention as implying desire, I am skeptical, first of all, about the whole category of so-called motivated desires - desires that the agent himself generates by deliberating, forming intentions or otherwise. I do not deny the possibility of motivation arising entirely within intention-forming agency. I do not deny that desire may be aroused in this way. But I do not believe that desires are the sort of things that can be generated or arrived at. I believe that desires, like feelings or emotions, are the sort of things that we simply acquire naturally or circumstantially. Secondly, I disagree that intention implies desire in a placeholder sense. My view is that desire, as an identifiable mental attitude, normally precedes and grounds the phenomenon of intending and that intention sublates desire. That is, in order for us to actively form an intention about an object, an action or a state of affairs, not only must it be present in our consciousness, but also it must arouse in us a desire or a desire-like state - an appetite, a feeling, an emotion, an inclination, a striving or a longing.

Not every object of our consciousness arouses a desire or an aversion in us. Some objects don't have an affective or appetitive impact on us; they leave us cold. But some objects arouse a positive or a negative inclination in us. It is the arousal of such inclinations or disinclinations by objects that we perceive, remember, imagine or represent in some way that sets the stage for the active formation of intentions. By why is the presence of a felt (dis)inclination a necessary factor in active intention-formation? It is necessary because whatever is intended is intended under the guise of the good, and, as I argued in the previous chapter, value or disvalue is primarily apprehended in and through feelings, emotions or desires. Intending is existentially committing oneself to bringing about a valuable state of affairs or making it one's business to bring it about. Valuable states of affairs, or potential projects, are given or presented to us first in our conscious affective states, and these states are what sets our faculties of reason and will in motion. A felt (dis)inclination is a necessary factor in active intention-formation because willing or intending requires and is based on value perceptions or motives. Our feelings, desires or desire-like states - our hunger, our thirst, our fear of pain, our love of music or literature, our sympathy, etc. - all

present themselves to our reason and will in the form of potential motives or purposes of action⁶. These states are real and simply happen to us - they are acquired naturally or circumstantially.

Finally, my view that desire as an identifiable psychological state precedes and grounds the event of forming an intention might face the objection that it cannot account for the sense in which it is possible for a person to do intentionally something that he doesn't want or desire to do. We often hear people say that they intentionally did something that they didn't want (or desire) to do. But can someone really intentionally do something that he doesn't desire to do? There is no natural reason to suppose not. If someone can act intentionally against his reason (or rational judgement), it must be possible for him to act intentionally against his desire too. This might sound puzzling because, intuitively, it follows from the fact A X-ed intentionally that A wanted to X.

To make sense of this puzzle, it is often suggested, we need to identify the sense in which we want (or desire) to do everything that we do intentionally, on the one hand, and the sense in which we can do something intentionally without wanting (or desiring) to do it. That is, there must be one sense of 'want' or 'desire' in which it is impossible to do something intentionally without wanting to do it, and another sense in which it is possible to do something intentionally without wanting or desiring to do it. Or there must be a kind of want or desire such that it is impossible to do something intentionally without having that want or desire, and another kind such that it is possible to do something intentionally without having that want or desire. And it is usually proposed that the first kind of desire is a motivated desire, and the second kind an unmotivated desire. This distinction does not, however, seem to be necessary for a plausible account of how one can intentionally A without wanting or desiring to A. The notion that intention sublates (i.e. depends on and transcends by affirmation or negation) desire makes room for this possibility as it does for the phenomenon of rationality in akratic actions.

⁶ This account of the relation of intention and desire draws on Pfänder's (1967) treatment of the phenomenology of willing and motivation, but departs from it in many crucial respects. One such respect is the central issue of whether desire is a necessary antecedent or ingredient of the phenomenon of willing or intending. Pfänder thinks that it is not, arguing that 'there are acts of willing in whose performance neither [desires] toward nor [aversions] against the projects in question can be discovered [- as when,] after completely calm and reasonable consideration and insight, one decides for or against something in accordance with this insight' (p. 25). My view that the will is an affective capacity and that intention sublates desire is clearly in disagreement here.

Conclusion

This thesis has attempted to investigate the problem of the constitution of human action, or the problem of how what we do depends on what we believe and what we want, or how our actions depend on our reason and our will. Because the problem of the constitution of human action is particularly unwieldy, there are a number of important aspects of it that either I did not touch at all or I touched but superficially, and that would therefore benefit from further investigation. By way of a conclusion, I would like to flag up just three such aspects or issues.

The first is the nature of voluntariness as a dimension of human agency, on the one hand, and its relation to intentionality, on the other. In Chapter One, I developed a concept of human action as intentional action by a self-governing agent. Here, I argued, *inter alia*, that an intentional action is an action done for reasons, that doing something intentionally involves intending to do it, and that there is no principled distinction between intentional action and self-governed action. I did not include the concept of voluntariness in the development of my notion of human action as an intentional action by a self-governing agent. Recently, Hyman (2015) has argued for the existence of four irreducibly different dimensions of human agency, with voluntariness as one of them. I would like to suggest that further research on the problem of the constitution of human action look into the relation between voluntariness and intentionality.

The second is the nature and intentionality of spontaneous or automatic actions. In Chapters Two and Three, I raised objections to Bratman's and Velleman's theories of human agency on the ground that they had no resources to account for the intentionality of spontaneous actions. In Chapter Two, I briefly attempted to account for it by arguing that the intentionality of our automatic or spontaneous actions is, at least partly, more plausibly explained by the habitual inclinations or willingness with which individuals antecedently are disposed to choose or pursue the objects of those actions. Once you are habitually inclined or willing to do something, you are normally led to do it without reflecting or forming an intention. I believe that the subject of the intentionality of spontaneous actions is more complex than my rather brief treatment might suggest, and so deserves further investigative attention.

The third and last issue for further research is the significance of agency for metaethics. There has recently emerged a metaethical position, known as constitutivism, according to which the norms of morality and their authority arise out of the nature of agency. My thesis investigated the problem of the constitution of human action as intentional action by a self-governing agent. Intentional action is presumably the standard form of human behaviour. Minimally, we do something intentionally when we do it by our own will. As such, intentional action is the action of the person as a whole, and not of some sub-personal mechanism or some psychological subsystem operating independently inside the person. And because they are actions of the whole person, intentional actions are behaviours that are typically amenable to certain characteristic sorts of evaluations (e.g. social, legal, prudential or moral), or behaviours for which a person is answerable in distinctive ways. This implication of intentionality suggests that human action is subject to, or governed by, certain norms. Indeed, we often wonder what we have reason (or ought) to do under our present circumstances, or whether there is a reason to be in a particular mental state (worry, fear, anger or joy) that we or others are in; and we generally assume that there are things that people have no reason to do under any circumstances. But what is the source of our reasons to act (or be) one way rather than another? And what makes an instance of human behaviour an intentional action? I initially wanted to investigate both of these questions in my thesis, but, for some reasons, I couldn't. I addressed only the second. Future research on the problem of the constitution of human action might look into the issue of whether there is any significant connection between the intentional and normative dimensions of human action.

Bibliography

- Aguilar, J., Buckareff, A. & Frankish, K. (2011). 'Introduction'. In Aguilar, J., Buckareff, A. & Frankish, K. (Eds.), *New Waves in Philosophy of Action*. (pp. 1-9). Palgrave Macmillan.
- Ajzen, I. (1985). 'From Intentions to Actions: A Theory of Planned Behavior'. In Kuhl, J. & Beckman, J. (Eds.), *Action Control: From Cognition to Behavior*. (pp. 11-39). Springer-Verlag.
- Alvarez, M. (2009). 'Acting Intentionally and Acting for a Reason'. *Inquiry*, 52(3), 293-305.
- Anscombe, G. E. M. (1963). *Intention* (2nd ed.). Cornell University Press.
- Aquinas, T. (1969). *Summa Theologiae*. Volume 16. (Trans. Thomas Gilby). Cambridge University Press.
- Aristotle (2009). *The Nicomachean Ethics* (Trans. David Ross). Oxford University Press.
- Arpaly, N. (2021). 'Three Dogmas of Agency Theory'. In Chang, R. & Sylvan, K. (Eds.), *The Routledge Handbook of Practical Reason*. (pp. 196-213). Routledge.
- Audi, R. (1993). *Action, Intention, and Reason*. Cornell University Press.
- Audi, R. (1986). 'Acting for Reasons'. *The Philosophical Review*, 95(4), 511-546.
- Baars, B. (1992) 'The Many Uses of Error: Twelve Steps to a Unified Framework'. In B. Baars (Ed.), *Experimental Slips and Human Error: Exploring the Architecture of Volition*. (pp. 3-37). Springer Science.
- Badura, C. & Kind, A. (2021). 'Introduction: The Epistemic Role of Imagination'. In Badura, C. & Kind, A. (Eds.), *Epistemic Uses of Imagination*. Routledge.
- Baumeister, R., Gailliot, M. & Tice, D. (2009). 'Free Willpower: A Limited Resource Theory of Volition, Choice and Self-Regulation'. In Morsella, E., Bargh, J. & Gollwitzer, P. (Eds.), *Oxford Handbook of Human Action*. (pp. 487-508). Oxford University Press.
- Bergamin, J. (2017). 'Being-in-the-flow: Expert Copying as Beyond both Thought and Automaticity'. *Phenomenology and the Cognitive Sciences*, 16, 403-424.
- Boler, J. (1993). 'Transcending the Natural: Duns Scotus on the Two Affections of the Will'. *American Catholic Philosophical Quarterly*, 67(1), 109-126.
- Bonansea, B. (1965). 'Duns Scotus' Voluntarism'. In Ryan, K. & Bonansea, B. (Eds.), *John Duns Scotus, 1265-1965*. (pp. 83-121). The Catholic University of America.

- Boswell, P. (2018). 'Intelligibility and the Guise of the Good'. *Journal of Ethics and Social Philosophy*, 13(1), 1-31.
- Brand, M. (1984). *Intending and Acting: Toward a Naturalized Action Theory*. The MIT Press.
- Bratman, M. (2006). *Structures of Agency*. Oxford University Press.
- Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.
- Brook, A. & Wuerth, J. (2023). 'Kant's View of the Mind and Consciousness of Self'. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Retrieved from <<https://plato.stanford.edu/archives/spr2023/entries/kant-mind/>>.
- Campbell, L. (2015). *Action, Intention and Knowledge*. (Unpublished Doctoral Thesis). University of Cambridge.
- Carey, J. (2017). *Reason, Desire, and the Will: In Defense of a Tripartite Moral Psychology*. (Unpublished Doctoral Thesis). University of California, Berkeley.
- Chartrand, T. & Bargh, J. (2002). 'Nonconscious Motivations: Their Activation, Operation, and Consequences'. In Tesser, A., Stapel, A. & Wood, J. (Eds.), *Self and Motivation: Emerging Psychological Perspectives*. American Psychological Association.
- Clark, P. (2010). 'Aspects, Guises, Species, and Knowing Something to Be Good'. In Tenenbaum, S. (Ed.), *Desire, Practical Reason, and the Good*. (pp.234-244). Oxford University Press.
- Crittenden, P. (2012). *Reason, Will and Emotion: Defending the Greek Tradition Against Triune Consciousness*. Palgrave Macmillan.
- Davidson, D. (2001). *Essays on Actions and Events* (2nd ed.). Clarendon Press.
- Davis, W. (2010). 'The Causal Theory of Action'. In O'Connor, T. & Sandis, C. (Eds.), *A Companion to the Philosophy of Action*. (pp. 32-39). Wiley-Blackwell.
- Di Nucci, E. (2013). *Mind Out Of Action: The Intentionality of Automatic Actions*. VDM Verlag Dr. Müller.
- Finnis, J. (1998). *Aquinas: Moral, Political, and Legal Theory*. Oxford University Press.
- Flanagan, J. (1997). *Quest for Self-Knowledge: An Essay in Lonergan's Philosophy*. University of Toronto Press.
- Ford, A. (2018). 'The Province of Human Agency'. *Noûs*, 52(3), 697-720.
- Frankfurt, H. (2006). *Taking Ourselves Seriously & Taking It Right* (Ed. D. Satz). Stanford University Press.
- Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge University Press.

- Fried, I., Haggard, P., He, B. & Schurger, A. (2017). 'Volition and Action in the Human Brain: Processes, Pathologies, and Reasons'. *The Journal of Neuroscience*, 37(45), 10842-10847.
- Gawande, A. (2009, March 23). 'Hellhole'. *The New Yorker*. <<https://www.newyorker.com/magazine/2009/03/30/hellhole>>.
- Goldman, A. (1970). *A Theory of Human Action*. Prentice-Hall.
- Grahek, N. (2001). *Feeling Pain and Being in Pain*. The MIT Press.
- Hampshire, S. & Hart, H. L. A. (1958). 'Decision, Intention, and Certainty'. *Mind*, 67, 1-12.
- Harman, G. (1986). 'Willing and Intending'. In Grandy, R. & Warner, R. (Eds.), *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. (pp. 363-380). Clarendon Press.
- Harman, G. (1976). 'Practical Reasoning'. *The Review of Metaphysics*, 29(3), 431-463.
- Hassin, R., Aarts, H., Eitam, B., & Custers, R. (2008). In Morsella, E., Bargh, J. & Gollwitzer, P. (Eds.), *Oxford Handbook of Human Action*. (pp. 549-566). Oxford University Press.
- Hegel, G. W. F. (2010). *The Science of Logic* (trans. & ed. G. Di Giovanni). Cambridge University Press.
- Helminiak, D. (2021). 'A Genuinely Scientific Psychology Based on Lonergan's Analysis of Consciousness'. *Clinical Psychiatry*, 7(S3:92), 1-4.
- Helminiak, D. (2014). 'More Than Awareness: Bernard Lonergan's Multi-Faceted Account of Consciousness'. *American Psychological Association*, 34(2), 116-132.
- Herberstein, M. and Wignall, A. (2011). 'Introduction: Spider Biology'. In M. Herberstein (Ed.), *Spider Behaviour: Flexibility and Versatility*. (pp. 1-30). Cambridge University Press.
- Heuer, U. (2014). 'Intentions and the Reasons for Which We Act'. *Proceedings of the Aristotelian Society*, 114 (Part 3), 291-315.
- Hieronymi, P. (2022). 'What is a Will?'. In Maoz, U. & Sinnott-Armstrong, W. (Eds.), *Free Will: Philosophers and Neuroscientists in Conversation*. Oxford University Press.
- Hieronymi, P. (2009). 'The Will as Reason'. *Philosophical Perspectives*, 23, 201-220.
- Hobart, R. E. (1934). 'Free Will as Involving Determination and Inconceivable Without It'. *Mind*, 18(169), 1-27.
- Hobbes, T. (1996). *Leviathan*. Oxford University Press.
- Holton, R. (2009). *Willing, Wanting, Waiting*. Clarendon Press.
- Hume, D. (1998). *An Enquiry Concerning the Principles of Morals*. Clarendon Press.
- Hume, D. (1960). *A Treatise of Human Nature*. Clarendon Press.

- Hursthouse, R. (1991). 'Arational Actions'. *The Journal of Philosophy*, 88(2), 57-68.
- Hyman, J. (2015). *Action, Knowledge, and Will*. Oxford University Press.
- James, W. (1981) [1890]. *The Principles of Psychology*. Harvard University Press.
- Jian, J. (2023). 'Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)'. In *Australasian Journal of Philosophy*, 101(2), 360-374.
- Jones, L. (1988). 'Motor Illusions: What Do They Reveal About Proprioception?'. *Psychological Bulletin*, 103 (1), 72-86.
- Kant, I. (2015). *Critique of Practical Reason*. Cambridge University Press.
- Kant, I. (1998). *Critique of Pure Reason*. Cambridge University Press.
- Kant, I. (1997). *Groundwork of the Metaphysics of Morals*. Cambridge University Press.
- Katsafanas, P. (2013). *Agency and the Foundations of Ethics: Nietzschean Constitutivism*. Oxford University Press.
- Kim, J. (2011). *Philosophy of Mind* (Third Edition). Westview Press.
- Korsgaard, C. (2014). 'The Normative Constitution of Agency'. In Vargas, M. & Yaffe, G. (Eds.), *Rational and Social Agency: The Philosophy of Michael Bratman*. (pp. 190-214). Oxford University Press.
- Korsgaard, C. (2011). 'Self-Constitution and Irony'. In Lear, J., *A Case for Irony*. Harvard University Press.
- Korsgaard, C. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press
- Korsgaard, C. (2008). *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford University Press.
- Korsgaard, C. (2006). 'Morality and the Logic of Caring'. In Frankfurt, H., *Taking Ourselves Seriously & Taking It Right* (Ed. D. Satz). (pp. 55-76). Stanford University Press.
- Korsgaard, C. (1996b). *The Sources of Normativity*. Cambridge University Press.
- Korsgaard, C. (1996a). *Creating the Kingdom of Ends*. Cambridge University Press.
- Kuhl, J. (1987). 'Action Control: The Maintenance of Motivational States'. In Halisch, F. & Kuhl, J. (Eds.), *Motivation, Intention and Volition*. (pp. 279 -291). Springer-Verlag.
- Kusch, M. (2024). 'Psychologism'. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Retrieved from <<https://plato.stanford.edu/archives/spr2024/entries/psychologism/>>.

- Lage, C., Wolmarans, D. W., & Mograbi, D. (2022). 'An Evolutionary View of Self-Awareness'. *Behavioural Processes*, 194, 104543.
- Langton, R. (2004). 'Intention as Faith'. In Hyman, J. & Steward, H. (Eds.), *Agency and Action*. (pp. 243-258). Cambridge University Press.
- Lecky, P. (1945). *Self-Consistency: A Theory of Personality*. Island Press.
- Lee, S. (1998). 'Scotus on the Will: The Rational Power and the Dual Affections'. *Vivarium*, 36(1), 40-54.
- Levy, N. & Bayne, T. (2004). 'A Will of One's Own: Consciousness, Control, and Character'. *International Journal of Law and Psychiatry*, 27, 459-470.
- Lewin, K. (1951). 'Intention, Will and Need'. In Rapaport, D. (Ed.), *Organization and Pathology of Thought: Selected Sources*. (pp. 95-153). Columbia University Press.
- Lombardi, N. (2013). *The Father's Will: Christ's Crucifixion and the Goodness of God*. Oxford University Press.
- Lonergan, B. (2017) [1972]. *Method in Theology*. University of Toronto Press.
- Lonergan, B. (2016). *A Second Collection*. University of Toronto Press.
- Lonergan, B. (2004). *Philosophical and Theological Papers 1965-1980*. University of Toronto Press.
- Lonergan, B. (2002). *The Ontological and Psychological Constitution of Christ* (Trans. M. Shields). University of Toronto Press.
- Lonergan, B. (1992) [1957]. *Insight*. University of Toronto Press.
- MacIntyre, A. (1986). 'The Intelligibility of Action'. In Krausz, M. Burian, R. & Margolis, J. (Eds.), *Rationality, Relativism, and the Human Sciences*. M. Nijhoff.
- McCann, H. (1991). 'Settled Objectives and Rational Constraints'. *American Philosophical Quarterly*, 28(1), 25-36.
- Malle, B. & Knobe, J. (1997). 'The Folk Concept of Intentionality'. *Journal of Experimental Social Psychology*, 33, 101-121.
- Mele, A., Nadelhoffer, T. & Khoudary, M. (2021). 'Folk Psychology and Proximal Intentions'. *Philosophical Psychology*, 34(6), 761-783.
- Mele, A. (2019). 'On Snubbing Proximal Intentions'. *Philosophical Studies*, 176, 2833-2853.
- Mele, A. (1992b). 'Acting for Reasons and Acting Intentionally'. *Pacific Philosophical Quarterly*, 73, 355-374.

- Mele, A. (1992a). *The Springs of Action: Understanding Intentional Behavior*. University of Oxford Press.
- Millgram, E. (2001). 'Practical Reasoning: The Current State of Play'. In Millgram, E. (Ed.), *Varieties of Practical Reasoning*. (pp. 1-26). The MIT Press.
- Miller, G., Galanter, E. & Pribram, K. (1960). *Plans and the Structure of Behavior*. Holt, Rinehart and Winston, Inc.
- Nabert, J. (1994). *L'Expérience Intérieure de la Liberté et Autres Essais de Philosophie Morale*. Presses Universitaires de France.
- Nagel, T. (1970). *The Possibility of Altruism*. Princeton University Press.
- New, C. (1999). *Philosophy of Literature: An Introduction*. Routledge.
- O'Brien, L. (2007). *Self-Knowing Agents*. University of Oxford Press.
- Orsi, F. (2015). 'The Guise of the Good'. *Philosophy Compass*, 10(10), 714-724.
- O'Shaughnessy, B. (2008). *The Will* (vol. 2). Cambridge University Press.
- O'Shaughnessy, B. (2004). 'Theories of Bodily Will'. In T. Pink & M. Stone (Eds.), *The Will and Human Action: From Antiquity to the Present Day*. (pp. 197-211). Routledge.
- O'Shaughnessy, B. (1991). 'Searle's Theory of Action'. In Lepore, E. & Van Gulick, R. (Eds.), *John Searle and His Critics*. (pp. 271-287). Blackwell.
- Pereboom, D. (2006). 'Kant on Transcendental Freedom'. *Philosophy and Phenomenological Research*, 73(3), 537-567.
- Pfäander, A. (1967). *Phenomenology of Willing and Motivation and Other Phenomenologica* (trans. H. Spiegelberg). Northwestern University Press.
- Price, A. W. (2004). 'Aristotle, the Stoics and the Will'. In T. Pink & M. Stone (Eds.), *The Will and Human Action: From Antiquity to the Present Day*. (pp. 29-52). Routledge.
- Proust, J. (2004). *La Nature de la Volonté*. Gallimard.
- Proust, J. (2019). *La Volonté et Ses Pathologies: Psychologie Expérimentale et Théorie de l'Âme chez Théodule Ribot*. (Unpublished Doctoral Thesis). Université Paris 1 Panthéon Sorbonne.
- Quinn, W. (1993). 'Putting Rationality in Its Place'. In Morris, C. & Frey, R. (Eds.), *Value, Welfare, and Morality*. Cambridge University Press.
- Raz, J. (2022). *The Roots of Normativity*. Oxford University Press.
- Raz, J. (2011). *From Normativity to Responsibility*. Oxford University Press.
- Raz, J. (1999). *Engaging Reason: On the Theory of Value and Action*. Oxford University Press.

- Ricoeur, P. (2009). *Philosophie de la Volonté. Le Volontaire et l'Involontaire*. Points.
- Robberechts, L. (1964). 'Quelques Théories de la Liberté. Autour de Jean Nabert'. *Revue Philosophique de Louvain*, 62(74), 233-257.
- Ross, W. D. (1924). *Aristotle's Metaphysics* (vol. 1). Clarendon Press.
- Saemi, A. (2017). 'Self-Knowledge and the Guise of the Good'. *Analytic Philosophy*, 0(0), 1-10.
- Saemi, A. (2014). 'The Guise of the Good and the Problem of Over-Intellectualism'. *Journal of Value Inquiry*, 48(3), 489-501.
- Saint Augustine (1998). *Confessions*. Oxford University Press.
- Schroeter, F. (2004). 'Endorsement and Autonomous Agency'. *Philosophy and Phenomenological Research*, 69(3), 633-659.
- Schueler, G. F. (1995). *Desire: Its Role in Practical Reason and the Explanation of Action*. MIT Press.
- Schwenkler, J. (2024). 'How Temptation Works'. *Journal of Ethics. And Social Philosophy*, 27(3), 443-468.
- Scotus, J. D. (2017). *Selected Writings on Ethics* (Ed. & trans. Thomas Williams). Oxford University Press.
- Scotus, J. D. (1986). *On the Will and Morality* (Select. & trans. Allan B. Wolter). Catholic University of America Press.
- Searle, J. (2001). *Rationality in Action*. The MIT Press.
- Searle, J. (1991). 'Response: The Background of Intentionality and Action'. In Lepore, E. & Van Gulick, R. (Eds.), *John Searle and His Critics*. (pp. 289-299). Blackwell.
- Setiya, K. (2007). *Reasons Without Rationalism*. Princeton University Press.
- Setiya, K. (2003). 'Explaining Action'. *The Philosophical Review*, 112(3), 339-393.
- Shepherd, J. (2015). 'Conscious Control Over Action'. *Mind & Language*, 38(3), 320-344.
- Siclari, F., Giulio, T., & Bassetti, C. (2012). 'Death by Sleepwalker'. *Scientific American Mind*, 23(3), 38-41.
- Smith, M. (2014). 'Scanlon on Desire and the Explanation of Action'. In R. J. Wallace, R. Kumar & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*. (pp. 79- 97). Oxford University Press.
- Steele, C. (2004). 'The Effect of the Will on Judgement: Thomas Aquinas on Faith and Prudence'. In T. Pink & M. Stone (Eds.), *The Will and Human Action: From Antiquity to the Present Day*. (pp. 78-98). Routledge.

- Steenburg, D. J. (1994). *Nature and History in the Knowledge of Value: A Study in Bernard Lonergan's Account of Value*. (Unpublished Doctoral Thesis). McMaster University.
- Stern, R. (2021). "How is Human Freedom Compatible with the Authority of the Good?" Murdoch on Moral Agency, Freedom, and Imagination'. *Proceedings of the Aristotelian Society*, 122 (Part 1), 1-26.
- Stout, R. (2005). *Action*. Acumen.
- Stocker, M. (1979). 'Desiring the Bad'. *Journal of Philosophy*, 76(2), 738-753.
- Tenenbaum, S. (2007). *Appearances of the Good*. Cambridge University Press.
- Vallacher, R. & Wegner, D. (2011). 'Act Identification Theory'. In Van Lange, P., Kruglanski, A. & Higgins, E. (Eds.), *Handbook of Theories of Social Psychology* (Vol. 1). (pp. 327-348). SAGE.
- Vallacher, R. & Wegner, D. (1987). 'What Do People Think They're Doing? Act Identification and Human Behavior'. *Psychological Review*, 94(1), 3-15.
- Velleman, J. D. (2007b). 'What Good is a Will?'. In Leist, A. (Ed.), *Action in Context*. (pp. 193-215). De Gruyter.
- Velleman, J. D. (2007a). 'Introduction to the David Hume Edition Series'. In Velleman, J. D., *Practical Reflection*. (pp. xi-xxxix). CSLI Publications.
- Velleman, J. D. (2006). *Self to Self: Selected Essays*. Cambridge University Press.
- Velleman, J. D. (2000). *The Possibility of Practical Reason*. Oxford University Press.
- Velleman, J. D. (1999). *Practical Reflection* (Second Edition). Princeton University Press.
- Vallacher, R. & Wegner, D. (1985). *A Theory of Action Identification*. Taylor & Francis.
- Wallace, R. J. & Kieseewetter, B. (2024). 'Practical Reason'. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Retrieved from <<https://plato.stanford.edu/archives/fall2024/entries/practical-reason/>>.
- Wallace, R. J. (2006). *Normativity and the Will: Selected Papers on Moral Psychology and Practical Reason*. Clarendon Press.
- Watson, G. (2004). *Agency and Responsibility: Selected Essays*. Clarendon Press.
- Watson, G. (2003). 'The Work of the Will'. In Stroud, S. & Tappolet, C. (Ed.), *Weakness of the Will and Practical Irrationality*. (pp. 172-200). Clarendon Press.
- Wegner, D., Vallacher, R., Kiersted, G. & Dizadi, D. (1986). 'Action Identification in the Emergence of Social Behavior'. *Social Cognition*, 4, 18-38.
- Wiland, E. (2012). *Reasons*. Continuum.

Wittgenstein, L. (1958). *Philosophical Investigations*. Basil Blackwell.

Zhu, J. (2004). 'Intention and Volition'. *Canadian Journal of Philosophy*, 34(2), 175-193.