Philosophy of Action: A Contemporary Introduction

Sarah K. Paul

Under contract with Routledge Press

Table of Contents

Acknowledgements

When I first arrived at graduate school to study philosophy, I didn't know what the philosophy of action was. I certainly didn't go with the intention of making any kind of intensive study of it. That I ended up becoming captivated by the topic and writing a dissertation on it is entirely due to the boundless enthusiasm and inexhaustible patience of Michael Bratman. These debates came alive when I saw them through his eyes, and as did the idea that I might one day be able to contribute something to the discussion. Much of what is in this book I learned from Michael, though he cannot be blamed for my errors and oversights. I am profoundly grateful for his support over the years, and fundamentally this book is for him.

I was subsequently welcomed into the community of philosophers of action by many people, but there are some I'd like specially to thank. Kieran Setiya, Luca Ferrero, Sergio Tenenbaum, and Sarah Buss all in various ways made me feel included and part of the conversation. Matthias Haase invited me to be a part of a "Netzwerk" of people interested in agency from all over Europe as well as the U.S. Participating in those meetings introduced me to a lot of wonderful people and greatly broadened my understanding of the variety of approaches one could take to these questions. And Matty Silverstein did the same by including me in his terrific series of workshops on "Normativity and Practical Reasoning."

My first tenure-track job was at the University of Wisconsin—Madison, and I owe that institution and my dear colleagues there a great deal. I agreed to write this book largely because Larry Shapiro told me to, and as my mentor, his thoughtful advice never erred. As my friend, I don't know about the advice, but Larry and Steve Nadler tolerantly let me air my frustrations with the writing process during our long runs around the arboretum. Elliott Sober, Russ Shafer-Landau, Mike Titelbaum, Alan Sidelle, and James Messina were also sources of great help and support. Much of this book was written with the aid of very generous research funding from the University of Wisconsin, including the Vilas Associates Award and Vilas Early Career Investigator Award.

Many thanks are also due to the students who took my undergraduate courses and graduate seminars in the philosophy of action over the years. There are too many to mention, but their creative examples and acute observations infuse what I have written here. Special thanks to Megan Fritts, Camila Hernandez Flowerman, and Ben Schwan for help with some of the research that contributed to this book.

The chapter on self-control was written with the support of a grant from the Philosophy and Science of Self-Control project directed by Al Mele and funded by the Templeton Foundation. I learned a great deal about the topic from the other researchers who were part of that project, and I'm grateful to Al for the support. The grant was for work done in collaboration with Jennifer Morton, who has been my dearest friend, sounding-board, advocate, and critic since we met in graduate school many years ago. The book was finally finished while at New York University – Abu Dhabi, and I'm very grateful for the generous research funding and great colleagues that I've found here.

Thanks also to my parents, Lowell and Sheila, and my husband Bas. Bas is a true partner and co-parent without whose help and perspective I couldn't do much of anything. I can't say my daughter Ivy specifically helped much with finishing this book, but the delay was completely worth it. I'm grateful to Aneena Noor, without whose help with childcare I could not have finished this book during a pandemic. Thanks also to my editor, Andy Beck, for his patience with me when the book took much longer to write than I thought it would.

Finally, for specific comments on the manuscript, I'm deeply indebted to David Velleman, Kieran Setiya, Sergio Tenenbaum, Luca Ferrero, Matty Silverstein, Mikayla Kelly, Michael Bratman, Jennifer Morton, Larry Shapiro, Sarah Buss, and (especially) an anonymous reader for Routledge Press. This book is far better because of their generosity.

1. Introduction: What is the Philosophy of Action?

In a bizarre incident that became the subject matter of *Palsgraf v. Long Island Railroad Co.*, a man is running to catch a departing train. A platform guard shoves him toward the train from behind while a member of the train's crew pulls him into the car. In the process, the man drops the package he is carrying. The package turns out to contain fireworks, which explode and cause a heavy scale ten feet away to topple over onto Helen Palsgraf, a woman who is standing on the platform. She suffers injuries as a result.

This case involves a number of people doing things, both intentionally and by accident. To understand what happened, as the court must try to do in adjudicating the suit, a variety of questions must be answered. Why did the platform guard shove the traveler? Was his intention to help him catch the train, or did he do it out of malice? Did the traveler drop the package on purpose, expecting it to explode, or did he merely lose his grip on it? Is it right to say that injuring Palsgraf was something the platform guard did, or something the traveler did? Or was her injury nobody's doing, and simply the result of bad luck? The philosophy of action aims to understand precisely what questions like these are asking and how we might go about answering them.

To isolate exactly what we are interested in, it will be helpful to first explain what the topic is *not*, at least for the purposes of this book. First, the problem of action is not the same as the ethical problem of how we ought to act, or what we should be blamed and praised for. *Palsgraf* certainly raises questions about whether the guard and the traveler acted as they should have and whether they are to blame for the injuries they caused. These, however, are separate subject matters. Our first task is to identify what it is we are referring to when we ask whether someone *acted* ethically, or what she ought to *do*. As G.E.M. Anscombe famously wrote, "…it is not profitable for us at present to do moral philosophy; that should be laid aside at any rate until we have an adequate philosophy of psychology …" (1958, 1). It may turn out in the end that ethics and the philosophy of action are intertwined, perhaps because rational agency consists in pursuing the Good or in doing what one takes to be morally required. But such insights should be the conclusion of our investigation and not premises in it.

Second, Anscombe's admonition that we do philosophy of psychology might wrongly suggest that we are interested in the so-called "mind-body problem" or the problem of mental causation. These labels refer to thorny issues about how to understand the relationship between the mind, revealed to us from within as consisting of conscious thoughts and experiences, and the physical body, as revealed to us by empirical investigation. How could the conscious experience of seeing orange be nothing more than some neural activity in the brain? And how could it not be, if having such experiences can have bodily consequences like uttering the sentence "What a beautiful shade of orange!" More generally, it seems that we can give physical explanations and mental explanations of the very same bodily event. When the guard's arm stretches out toward the traveler, we can explain what happened in terms of neural signals and muscle contractions, or we can clarify that he wanted to help the traveler and believed that pushing him at that moment was the way to do it. How, if at all, can these two kinds of explanations be reconciled?

The answers to these questions will be a crucial part of the complete story of what happens when an embodied agent moves herself. But before we can assess whether there is actually a direct conflict between physical and mental explanations, we must answer the question of whether an agent's raising his arm is really the same thing as an arm's rising. That is, are actions identical to bodily movements, or are they in a different category altogether? If they are not identical, then perhaps the two explanatory frameworks are not in conflict. Further, even if we solved the problem of understanding how the mind is related to physical states of the world, we would still face the puzzle of locating agency within that psycho-physical activity. Knowing exactly what a desire is in physical terms, and how desires cause the body to move, does not answer the question of whether a desire's causing a bodily movement amounts to an agent acting. For that, we need to do some philosophy of action.

Third, the problem of action is not the same as the problem of free will. As traditionally conceived, that question concerns whether we can be said to act freely if determinism is true, in the sense of freedom that many philosophers take to be required for moral responsibility. Can we ever be held accountable for what we do if our actions are preordained, either by the physical laws of nature or by the omniscience of a divine being? If the guard freely chose to push the traveler, and is legitimately blameworthy or praiseworthy for his choice or deed, must it have been genuinely possible for him to have chosen instead not to push? And if so, how could the existence of this alternative path be consistent with the rest of what we understand about the universe?

These questions are certainly relevant to our conception of ourselves and other beings as agents, and there is nothing wrong with using the label "philosophy of action" in a way that includes philosophical work on free will. However, this book will largely set such questions aside. On the narrower conception of the topic that is our focus, the central puzzles are both prior to and independent of our investigation into free will. They are prior because the question of whether our actions are free, and whether alternative possible courses of action are ever available to us, presuppose a grasp of what an action *is*. They are independent because the interest is in what it is to act, whether or not it turns out that we have the capacity to act freely in the sense relevant to moral responsibility. In other words, even if we were to conclude that we have no free will, we would still be agents who spend our waking hours engaged in (unfree) actions of various kinds, and this would still be a phenomenon worth trying to understand.

Having trimmed back the hedges of surrounding issues, we are in a better position to think about the basic phenomenon of *doing* something. We will attempt to state the central questions more precisely in Chapter Two, but for now we can start with some intuitions. Think of all the things you have done just today. Perhaps you turned off your alarm, made coffee, spilled some on your shirt, arrived late to class, and spent some time fantasizing about lunch. "Doings" can be contrasted, first, with mere "happenings." Turning off your alarm was something that you did – you made it happen – but being awakened by it is something that happened to you. What exactly is the difference between making something happen and having something happen to you? Second, not everything you do is something you do intentionally, or as we more commonly say, "on purpose." Your spilling coffee on your shirt was presumably not something that you meant to do, whereas making the coffee was intentional. Similarly, while arriving to class and arriving late

to class seem to be one and the same event, we can assume that you intentionally went to class but did not intentionally arrive late. But what does it mean to do something intentionally or not, and how can the same event of your arriving to class at 9:23am be both an intentional action of yours and something you did unintentionally?

We assume that you intentionally went to class but did not intentionally arrive late because agents like us normally act in light of what we have good reason to do, or what it would make sense for us to do. There are many things to be said in favor of going to class, but not many in favor of being late. This suggests that whether or not your action was intentional is connected to whether you saw it as desirable, worth doing, or at least intelligible. That said, we do not always choose to do what we take to be the best thing to do. It seems perfectly possible to decide to sleep in and be late to class even though you sincerely believe that being punctual is more important than 15 more minutes of sleep. What on earth is going on when we intentionally do one thing while at the same time believing that another available option is substantially better? Another interesting feature of agents like us, and one that can be of use in combatting this kind of weak-willed action, is that we do not always simply make up our minds in the moment about what to do. We sometimes deliberate in advance and plan ahead for what to do later, thereby forming intentions for the future. "Tonight," you might think, "I'll set my alarm 15 minutes earlier so that I'll be on time to class tomorrow."

Some of our actions are mental – fantasizing about lunch – while others involve far-reaching consequences in the non-mental world. For that matter, some of the most consequential events we are involved in concern things that we intentionally *don't* do. Stanislav Petrov may have singlehandedly saved the world from all-out nuclear war when he refused to obey protocol and did not respond to what turned out to be a false alarm in the Soviet early warning system. And while singlehanded actions are often of particular interest to us, a great many of our actions are undertaken together with other people. We play dodgeball together, write books together, and participate in the institutions of society together. All these are the kinds of phenomena we will attempt to understand better in this book.

One way to motivate these puzzles about agency is to compare them to other well-known philosophical problems. For example, philosophy has long been occupied with investigating the nature of perception. We seek to better understand how the conscious mind is connected to the external world, and perception is one major way in which the mind interfaces with the world. We wrestle with questions like "How can we be in touch with objects that are external to our minds through the use of our senses?" and "Given the possibility of illusion, how can perception give us knowledge of reality?" Inspired by these questions, we might view action theory as a counterpart to the philosophy of perception. It is through acting that the mind interfaces with the world in the other direction, imposing itself on external reality rather than conforming itself to it. This framing will generate questions about action that tend toward the metaphysical, focusing on how agency fits together with the rest of the natural world.

A different contrast can be drawn between action and belief, thereby framing the topic in a way that will focus our attention on certain normative questions. Philosophers have long been interested in the question of why we should or shouldn't believe certain things – what counts as a

sufficient reason to believe? We attempt to understand what it is to reason well about what to believe (call this "theoretical" reasoning), so that we can avoid certain tempting forms of logical error. Think of Sherlock Holmes reasoning about who might have committed the murder and forming the belief that it was a one-legged man and his small accomplice, for the reason that the tracks in the dust indicate these unusual physical features. Viewed through this lens, the state or activity of believing is defined partly in terms of its being the conclusion of theoretical reasoning, or a way of responding to these kinds of reasons. Likewise, we act for reasons and engage in "practical" reasoning about what to do. Holmes might weigh the options of revealing his suspicions to the police or pursuing the one-legged man himself and decide to involve the police for the reason that they have the fastest boat. If action is the analogue to belief, we might similarly define it in part as a kind of response to our practical reasons, or as the conclusion of practical reasoning. From this perspective, action theory can be thought of as a counterpart to epistemology.

These are two somewhat different ways of situating action theory with respect to other areas of philosophy. We can motivate the inquiry even further by thinking about the practical relevance of the kinds of questions it promises to answer. Often, it is of great importance to us to know whether something that happened was an intentional action or not. In the *Palsgraf* case, much depended on whether the fireworks were set off intentionally or merely by accident. Other, related distinctions come up frequently in legal settings. The criminal law generally requires that there be a "voluntary act" in order for a crime to have occurred, for example. Merely planning to rob a bank is usually not enough to incur criminal liability. In addition, it distinguishes between acts that are undertaken with intent, with knowledge, out of recklessness, and out of negligence. It treats certain crimes that are committed as a result of passion or provocation differently than those committed with premeditation, but it treats those done in ignorance of the law and those done with knowledge of their illegality as the same. These all seem, at least in part, to be an effort to distinguish between different kinds of action. Philosophers of action aim to determine whether such distinctions hold up under scrutiny, and to articulate the precise conditions under which they apply.

There is a rich history of philosophical investigation into agency and action, reaching back to the very beginning of philosophy as a discipline. However, action theory was not really considered an autonomous area of philosophical inquiry as distinguished from ethics, free will, and philosophy of mind until relatively recently. The groundbreaking work of Anscombe and Donald Davidson in the mid-to-late twentieth century defined the set of questions that frame contemporary research on agency. This book aims to introduce readers to these contemporary debates rather than to offer a detailed historical perspective.

In spite of being a fairly young field – or perhaps because of it – there is quite a lot of disagreement about where exactly the philosophy of action should start and where it aims to end up. The organization of the material here reflects my perspective on the best way to progress through the topic, but this is one perspective among many legitimate ones. Nearly all of the chapters have a serious claim to being the one that should go first, and for each of those, there are respectable philosophers who would argue that it should be left out altogether. The goal of Chapter Two is to lay out the thicket of overlapping questions, concepts, and phenomena that have been

used to frame the investigation of agency. Close attention to these different starting points will equip the reader to better understand what follows, since in my view, many apparently substantive disagreements can be traced back to divergences in this first step.

Chapter Three takes up the idea that the kind of action we should be interested in – rational or intentional action – is subject to a distinctive kind of *explanation*. When we ask why a person acted as he did, the answer can take the form of providing the person's reasons for so acting. Suppose we inquire into why Richie went to D.C. this weekend. If the explanation is that he boarded the wrong train, thinking it was going to Philadelphia, this reveals to us that he did not go to D.C. intentionally. 'Boarding the wrong train' is not a reason to go to D.C. But if the explanation is that he went to see a concert, the implication is that he did go there intentionally. One of the central debates in the field concerns the nature of this kind of rationalizing explanation. How exactly does citing a person's reason provide illumination of her action?

Chapter Four turns to the question of what an action *is*. We learn from Chapter Three that actions are the kind of thing that can be given a rationalizing explanation, but what kind of thing is that? This chapter delves into the metaphysical details. How are intentional actions related to unintentional ones? How are actions individuated from one another? Are some actions "basic," and if so, which ones? And most importantly, what is it that makes an event an action?

It is almost irresistible to think that the answer to this last question has something to do with what the agent of the action had in mind. Chapter Five therefore shifts the focus to the notion of intention. We commonly speak of agents as *having* intentions or *acting with* certain intentions. What do we mean by these locutions? We might choose not to take such ways of speaking literally, viewing them instead as mere figures of speech. But those who do take them literally generally understand the term 'intention' to refer to a kind of mental state or attitude. The challenge, then, is to understand what kind of attitude it is and how such attitudes are related to intentional action.

In a sense, Chapter Six starts our discussion over. Like Chapter Three, it explores a phenomenon that many philosophers take to be a criterion of the kind of action we should be interested in: "practical knowledge." When we act intentionally, we normally know what we are doing in a distinctively first-personal way. For you to know that I am currently writing a book about agency, you will have to use methods like perception, inference, or testimony. You might lean over to take a glance at my computer screen or ask my colleague what I am up to. But I can know that I am writing a book without waiting to observe what appears on the screen or performing any kind of conscious inference. Reflecting on the phenomenon of practical knowledge gives us the opportunity to step back and further assess some of the proposals we have seen so far. For instance, some theories of intention seem to offer better explanations of practical knowledge than others. What we make of this depends on how central to our topic we take practical knowledge to be.

According to one kind of view, practical knowledge is not only central to our agency but the goal or "constitutive aim" of acting. The proposal is that in addition to the various specific purposes we have for doing things, intentional action itself has a purpose, namely, to afford us knowledge of ourselves. Other approaches agree that action has a constitutive aim but deny that this aim is epistemic. One longstanding tradition holds that all action aims at doing what is best,

or at least what is good. Rival accounts claim that it aims at constituting or unifying the self, or at expressing the agent's drive to overcome resistance and expand her power. Against all of this, there are many who deny that action as such has any particular aim. Chapter Seven is dedicated to examining these debates.

Chapter Eight draws a distinction between mere intentional action and action "par excellence." The thought is that of all the things we go around doing intentionally or voluntarily, only some of them are truly autonomous. The latter actions represent what the agent truly wants, or who she truly is, or what she is fundamentally committed to. To make good on this idea, the challenge is to specify which elements or structures of the agent's psychology truly speak for her, such that the actions that are suitably related to those elements count as fully autonomous. Is it some special subset of her desires? Her values or character? The plans and policies she has set for herself?

The counterpart to fully autonomous action is weak-willed or "akratic" action. We act weakly when we succumb to temptation or otherwise choose not to do what we believe would be best. We procrastinate, over-indulge, cheat on our vows, and work too much. The weak-willed agent is not ignorant of the considerations that speak against her action, and she does not act from vice – it is not that she takes the worse action to be the better. Nor is she straightforwardly compelled to act as she does, in the sense of having no choice in the matter. Rather, as we might say, she simply fails to control herself. But how is it even possible to act both intentionally and weakly, and what would it mean to exercise self-control? This is the central puzzle of Chapter Nine.

Finally, Chapter Ten extends the focus to the case of acting together with other people. It is commonplace to speak of groups as deciding, intending, and doing various things. People play Scrabble, get married, elect presidents, and write laws together. Can we account for "shared" agency using only those resources mentioned in the first nine chapters? That is, can we understand it as a matter of individual agents with ordinary intentions who are each doing their part? Or do we need to introduce new notions like "group agents" or "group intentions?"

Though my own views on these issues will inevitably come through to some extent, I do not take myself to be arguing for any particular conclusions. The seminal texts are difficult enough that one cannot simply clarify what is going on in them without being somewhat opinionated. But given the divisions in the field, I will be satisfied if I can help to illuminate where different thinkers are disagreeing with one another and where they are talking past one another. My hope is that I have left sufficient room for readers to form their own views.

2. What Is the Problem of Action?

To begin our investigation into agency and action, we must first try to state more clearly and precisely what the central questions are. As we will see in subsequent chapters, there is a potentially bewildering array of different approaches to understanding what action is, or what it is to be an agent. This kind of vigorous debate tends to be the case in any area of philosophy, but the philosophy of action is particularly difficult in this regard. There is broad disagreement about the questions we should be interested in addressing and what our starting assumptions should be. To make matters worse, these disagreements are not always obvious or made explicit by the various parties in these debates, which can lead to theorists simply talking past one another.

The problem, I think, is that 'action' is a sub-class of a more general category that we might call 'activity' or 'behavior'. Any given philosophical investigation of action must begin with some idea of how this general category ought to be restricted. And there a variety of interests we might have in trying to draw these lines. Some think it obvious that our philosophical interest in action is broadly ethical, and so find ways of delimiting the topic that emphasize the connection to reason, responsibility, self-consciousness, and self-understanding. Others think the interesting puzzles are obviously metaphysical, and so frame the topic in a way that emphasizes the contrasts between actions and other kinds of events or occurrences in the world. Still others are interested in behaviors that are the manifestation of a certain kind of psychology, and so demarcate the topic in that way. The everyday notion of action is fluid enough to accommodate all of these purposes, and so we cannot simply rely on our intuitions about the meaning of the word to clarify what it is we are talking about. Yet for those of us who have spent significant time thinking about agency from a particular perspective, it is easy to forget that this is so (and I include myself in this).

This chapter will attempt to lay out the variety of ways that we might choose to frame the investigation, with the hope that we will be better able in subsequent chapters to understand why different theorists have ended up where they are.

1. Activity and passivity

The most general way to frame the central question is in terms of activity: what is it to *make something happen*? Harry Frankfurt seems to pose the puzzle in this way when he states that "The problem of action is to explicate the contrast between what an agent does and what merely happens to him" (1978, 157). Taken at face value, this is a very general metaphysical puzzle about what it is to be the source of change. Even chemicals and tornados are agents, understood in this way. It will be helpful here to introduce the somewhat antiquated term *patient* to serve as the antonym of 'agent'. If an agent is the source of some change, the patient is the thing that undergoes or suffers change – the thing that is acted upon.

Most philosophers of action do not take themselves to be investigating the nature of activity in this very broad and abstract sense that includes chemicals and tornados. Often, they introduce the term 'agent' or 'action' with some implicit restriction in mind. For instance, in Frankfurt's

initial remarks, only his use of the pronoun 'him' rather than 'it' makes clear that he is thinking specifically of human agents, or perhaps a somewhat wider class of living organisms (though he goes on to make this more explicit by rephrasing the question in terms of 'bodily movements'). To avoid confusion, it is important for philosophers of action to begin theorizing by making clear how we intend to restrict the class of agents and activities we are interested in, and importantly, to defend the choice to restrict the target in that particular way. Why is the kind of activity we have identified distinctive and philosophically interesting?

2. Goal-directedness

One possibility is to limit our focus to agency that is purposive or goal-directed. In philosophy, the phenomenon of goal-directedness is often referred to using the term *teleology*, from the Greek 'telos'. Whereas the acid does not have any purpose in corroding the metal, and the tornado does not have the goal of destroying the village, many actions do occur for the sake of some goal. This way of approaching the topic will still be quite inclusive, since most life-forms are capable of goal-directed self-change. For instance, the young sunflower adjusts itself throughout the day so that it continually faces the sun or other light source. It does this in order to get nourishment from light and because this movement facilitates pollination. What is distinctive about goal-directed self-movement is that it seems to involve *guidance*: the flower regulates its position with respect to the location of the light as it moves throughout the day. Similarly, the spider uses its capacity to produce silk for the purpose of building a web in a way that seems to be guided toward catching prey.

If we take the kind of purposive activity exhibited by sunflowers and spiders as well as human animals to be our starting-point, this will help us to avoid the assumption that agency must involve highly sophisticated cognitive capacities. Further, there is no doubt that teleology is a philosophically interesting phenomenon, and thus that there is a rationale for focusing specifically on what it is to be the source of a change that is directed at some goal or purpose. However, we might worry that this class of actions is still too broad. Arguably, the kinds of actions we are interested in must be performed by an entity that deserves to be called an *agent*. But goal-directed activity can take place without being attributable to an agent. For example, digestion is a process of change that is directed at the goal of converting food into energy, but we would not generally speak of the digestive system as an agent.

3. Attributability

In light of the foregoing concern, some approaches to framing the problem of agency emphasize *attributability*. On this way of thinking about it, it is essential to being an action that it is attributable to an agent, or to the person or organism as a whole, rather than to some sub-agential part of her or process within her. Though we might speak loosely of a person digesting her food, it is more accurate to say that her gastrointestinal system did the digesting. Similarly, when we attribute a person's behavior to instinct, habit, or to some event occurring in her brain, this is

generally a way of saying that it was not fully an action of *hers*. In this vein, Christine Korsgaard writes that "Unity is essential to agency … because an action, unlike other events whose causes in some way run through an agent, is supposed to be a movement, or an effecting of change, that is backed by the agent as a whole" (2014, 193). And Frankfurt asserts that a central problem in the philosophy of action is " … to specify when the guidance of behavior is attributable to an agent and not simply, as when a person's pupils dilate because the light fades, to some local process going on within the agent's body" (1977, 159).

One motivation for this is approach is that one might think attributability is necessary for, or deeply connected to, *accountability*. That is, in order to hold someone responsible for something that happened, it is plausible that we must first attribute what happened to him. If your professor is lecturing and hears someone make an objectionable remark, she cannot blame any particular person in the class until she identifies who it was that said it. Even then, the student might try to avoid responsibility by claiming 'it was a slip of the tongue', thereby attributing the apparent offensiveness of the remark to some part of him rather than conceding that he himself had said it. As mentioned in Chapter One, we should not assume at the outset that the assignment of moral responsibility perfectly tracks the contours of our agency. That said, there is clearly a deep connection between the two topics, and it is natural to think that this connection runs in part through the notion of attributability.

4. 'Actish' phenomenal quality

Yet another way of focusing our attention on the target is to think about the phenomenology of agency – the experiential aspect of performing an action. Some have maintained that there is a distinctive feeling involved when we act, such that any event that has this "actish phenomenal quality" is *ipso facto* an action. The phrase 'actish phenomenal quality' is owed to Carl Ginet, who also describes it as the "I-directly-make-it-happen phenomenal quality" (1990, 14). Others have characterized it as a feeling of authorship, ownership, or control. This feeling is present, it is claimed, when *you* move your arm and absent when someone else moves it for you. It is also absent in certain pathologies like Anarchic Hand Syndrome, in which people report that their hand moves itself in a goal-directed way – it might button up one's shirt, for example – but that they do not feel as though they themselves have moved it or that the resulting action is theirs. Focusing on the phenomenology of agency is a way of trying to home in on an intrinsic quality of action rather than referring to extrinsic qualities like its relationship to an agent or to some further goal.

5. Voluntary and intentional action

Historically speaking, the kind of action of interest to philosophers, theologians, and legal theorists has often been characterized as *voluntary* action (from the Latin *voluntas*, or will). The notion of voluntariness was originally tied directly to moral responsibility: the thought was that we

can only be held responsible for what we voluntarily do or allow to happen. Aristotle, for instance, introduces the notion in this way:

> "…it is only voluntary actions for which praise and blame are given; those that are involuntary are condoned, and sometimes even pitied. Hence it seems to be necessary for the student of ethics to define the difference between the Voluntary and the Involuntary…" (NE III 1109b 30-35)

Later, the notion came to be tied to the idea of *willing*. It became widely agreed-upon that in addition to the faculty of intellect, human beings also have the faculty of will which issues in volitions. The term 'voluntary' is thus often used more or less as a synonym for 'volitionally' or 'willingly'. However, it is unclear whether this latter understanding of voluntary action fits well the ethical origins of the concept. Aristotle observed that we tend to pardon acts that are the result of compulsion, and that they are to that extent involuntary. But it seems possible for the will to participate in a pardonable act, as when a person under torture finally tells his torturers what they want to know.

From the late nineteenth to the mid-twentieth century, the notion of the will fell from favor. It came to be seen as a mysterious and unscientific idea. As Friederich Nietzsche disparagingly wrote, "The 'inner world' is full of phantoms and will-o'-the-wisps; will is one of them" (1888, 135). Gilbert Ryle and Ludwig Wittgenstein similarly attacked the "myth" of volitions or acts of will as part and parcel of what they took to be a deeply mistaken hangover from the days of believing in an immaterial soul. On their view, to look for inner springs of action is simply to misunderstand the logic of concepts like 'voluntary'. These attacks on the idea of the will were broadly effective (though see Brian O'Shaughnessy's *The Will* for dissent); by the latter half of the twentieth century, the focus had largely shifted to "intentional" action, and that is where the field stands as of the writing of this book (more on this in a moment).

6. Rational action, or acting for reasons

A number of philosophers think that the focus should be restricted to rational action, or actions done for reasons. Many of the things we do are done in light of considerations that we take to justify the action: considerations in light of which the action is seen as desirable, ethical, fitting, instrumental to some further purpose, or otherwise rationally intelligible. One way to express this point is to say that this kind of action is subject to a distinctive kind of explanation we might call *rationalizing* explanation. As G.E.M. Anscombe puts it, intentional actions are those "to which a certain sense of the question 'Why?' is given application; the sense is of course that in which the answer, if positive, gives a reason for acting" (1963, 9) This way of approaching the problem of action focuses on what is distinctive about human agency, in contrast with at least most non-human animals. Whereas the sunflower and the spider are active in their goal-directed movements, which may be attributable to them and not merely to events inside them, most would deny that they act for reasons in the sense elicited by the 'Why?' question. They are simply not the kinds of creatures

who operate in what Wilfrid Sellars called "the logical space of reasons," where we traffic in the demand for justification. This approach also has the advantage of framing the question in a way that is clearly philosophical, whereas focusing on the difference between activity and passivity threatens to collapse into a predominantly empirical investigation.

7. Practical knowledge

Yet another way of circumscribing the domain of action theory appeals to a broader conception of the agent's first-personal perspective than mere phenomenological experience. This approach emphasizes the agent's awareness or knowledge not only *of acting*, but of *what* he is doing. In discussing the central importance of the 'Why?' question for understanding intentional action, Anscombe elaborates that there are at least two ways that it can be shown not to have application: 1) if the person replies "I was not aware that I was doing that," and 2) if she replies "I only knew that I was doing that by observing myself." To get the intuition, imagine that someone is stepping on your foot, and you say "What do you think you're doing?!" If the person sincerely replies, "Oh dear, I wasn't aware that that was your foot," or "Indeed, I see now that I am stepping on it!" then he cannot have been stepping on your foot intentionally. If he were, he would know that he was without needing to find out by looking. Anscombe further suggests that the knowledge we have of our intentional actions is not the product of inference. Thus, we can characterize the actions that should be the topic of philosophical theorizing as those things a person knows she is doing without inference or observation. We can call this distinctive kind of knowledge of our own actions "practical knowledge."

8. Intentional action

The term "intentional action" has cropped up a few times now. Most contemporary discussions of action actually claim that this is what they are giving an account of. The phrase is ripe for theorizing, since it does not have a precise or widely agreed-upon meaning. On the surface, it seems to be tightly connected to the idea of having an intention, though this is also a fairly obscure notion. Experimental work suggests that people use the adjective 'intentional' and the adverb 'intentionally' in fairly unsystematic ways. Though imperfectly, it seems to track many of the features that we are interested in – acting with purpose, knowingly, deliberately, for a reason, or in a way for which we can be held accountable. And sometimes, I suspect, it is used simply as a synonym for 'voluntary'. It is therefore useful as a way to get our intuitions pumping, though we must be careful not to put too much weight on those intuitions or on the word itself. Ultimately, our best philosophical theory of intentional action may not perfectly track the ordinary use of the word.

9. Intention

In addition to characterizing some actions as intentional, we also talk about what we intend to do, or the intention with which we are acting. You might form an intention in the morning to go to the party tonight, or you might arrive late at the party with the intention of seeming like a person with a full social calendar. Rather than beginning our theorizing with trying to understand what it is to do something intentionally, we might consider beginning with trying to understand what an intention is. For example, one might think that in order for an action to be intentional, it must specifically be intended by the agent or stand in some close relation to an intention the agent has. This would be to define intentional action in terms of some mental state or property, "intention." In a sense, this is a way of reviving and demystifying the obscure notion of volition with the hope that we can give an unproblematic and scientifically respectable account of intention.

10. Autonomy, identification, and self-governance

A further source of complication and confusion in the contemporary landscape of action theory is that some views focus primarily on explaining "self-governed" or "autonomous" action ('autonomy' comes from the Greek 'auto', which means 'self', and 'nomos' which means 'law'). The basic thought here is that some actions that are purposeful, voluntary, intentional, and known to the agent without observation or inference nevertheless fall short of being autonomous because the agent is alienated from them or otherwise fails to fully stand behind them. This might be the right way to describe weak-willed actions, for instance. Imagine an inveterate gambler who desperately wants to change her ways, but who still voluntarily, intentionally, and knowingly makes and carries out a plan to go to the track tonight. She exercises agency in going to the track, but there is a sense in which the ponies are governing what she does and not she herself. On this approach to theorizing about agency, the question is "what are the conditions under which the agent is fully governing herself, such that her actions 'speak for her'?" This can be confusing because the theories that are offered as answers to this narrower question about "full-blooded" agency or "agency par excellence" have sometimes mistakenly been taken to be views about the broader categories of voluntary or intentional agency.

11. Further choice points

a. Which cases are paradigmatic?

In any theoretical investigation, where we end up can depend significantly on where we start. The philosophy of action in particular proceeds in large part by reflecting on specific examples. But which examples should we treat as paradigmatic of the phenomenon we are interested in, and which should be treated as marginal or defective? Though it is rarely made explicit, disagreement about these questions is implicit in much of the seminal work on the topic. For instance, Davidson's examples tend to be couched in the past tense – something was done, or something happened – while Anscombe's examples are couched in the present – something is being done. And Michael Bratman gives priority to actions that are still in the future and the plans that

we make about them. As we will see, these apparently innocuous differences might matter quite a bit.

A second methodological difference concerns breakdowns in action – cases in which something goes wrong. Should we begin our investigation by taking into account examples of agents whose bodies fail to move as they will them to, or who tend to be subject to cognitive biases and other irrationalities, or who are deeply out of touch with reality? If we do this, then we are likely to end up with a theory of agency that is tailored to fit with such human flaws. On the other hand, we might think that theorizing about cases of defective action gets us off on the wrong foot. Rather, we should begin with ideal cases in which all of the agent's capacities are functioning optimally. We can then understand sub-optimal cases as defective versions of agential excellence. The latter approach is likely to generate a theory of action that raises questions about how often non-ideal human agents like ourselves manage to act at all.

b. Questions about action: conceptual or ontological?

Finally, we should aim to be clear about the kinds of philosophical questions we are asking when we investigate agency. When we ask what it is to act intentionally, or autonomously, or to be an agent, are we asking about the *concepts* 'agent' and 'action' – what we ordinarily mean when we call something an agent or an action? Or are we asking about the *referents* of these concepts – what something ordinarily is when we call it an action or an agent? Ultimately, we are surely interested in both the concepts and the reality, insofar as the two can come apart. But it is a recipe for deep confusion when philosophers blur these two levels of analysis or fail to specify which of them they are talking about. I think this kind of mistake occurs quite often in discussing the view known as the Causal Theory of Action, for example. The claims this kind of view makes about the connection between actions and causes tend to be at the level of ontology (i.e. at the level of what reality consists in), but they are sometimes taken up as if they are claims about concepts. Keeping track of this distinction will help quite a bit to clear up what people are disagreeing about.

Those who explicitly take themselves to be theorizing about the concept of action sometimes restrict their questions to what we might call the "agential standpoint" or the "practical perspective." That is, their project is to investigate the first-personal point of view of a being who is engaged in the activity of deliberating and determining for herself what she will do. This kind of approach denies that agential concepts are fundamentally empirical – that they get their content from playing a role in the individuation, explanation, and prediction of events. Other philosophers who do take agential concepts to be empirical in this way, or whose projects focus primarily on metaphysics rather than concepts, are often exasperated by the refusal of the former theorists to make contact with the empirical sciences.

12. Conclusion

In what follows, I will attempt where possible to flag which of these notions is at issue in the discussion. When a relatively neutral characterization is called for, I will generally use the term

'intentional action' to mean "the actions that are the primary topic of the philosophy of agency," whatever those turn out to be from a conceptual and ontological standpoint. This is the term that is most in vogue at the time of writing this book, though I have my worries about whether its ordinary meaning is robust enough to bear the weight that such theorizing puts upon it. Readers who share these worries are invited to substitute other ways of characterizing the target notion – 'autonomous action', for instance – and see whether the arguments work in the same way.

Suggested Reading

George Wilson and Sam Shpall provide a nice overview of the field in their entry under "Action" in the *Stanford Encyclopedia of Philosophy*. Harry Frankfurt discusses a number of the above distinctions – activity and passivity, goal-directedness, and attributability – in "The Problem of Human Action." Donald Davidson's papers "Agency and "Actions, Reasons, Causes" did much to define the central questions, as did G.E.M. Anscombe's monograph *Intention* (see especially sections 1-17). The "Introduction" to J. David Velleman's collection *The Possibility of Practical Reason* emphasizes the relevance of practical knowledge and helpfully illustrates the idea that an action might be more or less autonomous. The idea that action has a distinctive phenomenal quality can be found in Carl Ginet's *On Action* (see especially Chapter One), while Myrto Mylopolous and Joshua Shepherd offer a more general overview of agential phenomenology in "The Experience of Agency." And Thomas Pink and Martin Stone's *The Will and Human Action: From Antiquity to the Present Day* provides a very useful historical overview. Other good general resources include *Philosophy of Action* by Lilian O'Brien, *A Companion to the Philosophy of Action*, edited by Tim O'Connor and Constantine Sandis, and the *Routledge Handbook of the Philosophy of Agency*, edited by Luca Ferrero.

3. Action Explanation

Imagine that you walk into a room and see your friend Mariko peel a banana, put the peel into a blender, and throw the banana in the garbage. You ask her "Why are you doing *that*?" One possibility is that she might look in the blender and exclaim "Oops, that's not what I meant to do – I wasn't paying attention!" This answer seems to indicate that the action of blending the peel and discarding the banana was not intentional in the sense that we philosophers aim to understand. A second possibility is that she replies "I'm making a low-calorie, high-fiber smoothie! I want to get more fiber in my diet, and I read that all the calories are in the banana and all the fiber is in the peel." This explanation implies that Mariko is blending the peel intentionally. Whereas the first explanation merely identifies the cause of her happening to throw away a good banana – a lack of attention – the second one identifies a *reason* for throwing it away. Call this second kind of reply to the 'Why?' question a "rationalizing" explanation.

Rationalizing explanations of action reveal the light in which the agent viewed the action as worth doing, or something that it would make sense to do. In Donald Davidson's words, the explanation "… leads us to see something the agent saw, or thought he saw, in his action – some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable" (1963, 685). In this case, Mariko was making a banana-peel smoothie because she thought it would be healthy and slimming and therefore good, or at least intelligible. Of course, she may be mistaken about this; what matters for explaining her action is that she *thinks* there is a good reason for doing it. In contrast, although a lack of attention explains why she threw the banana away in the first case, it is not the kind of thing that could justify her doing so. Though we might say that inattentiveness is the "reason why" she threw it away, this sense of "reason" refers to a cause rather than a consideration that shows what happened to be good or rationally intelligible.

Action explanation may seem an odd place to start – should we not first inquire into what actions *are*, before we ask how the occurrence of such things can be explained? But as we saw in Chapter Two, many philosophers follow G.E.M. Anscombe in thinking that rationalizing explanations are central to understanding the kind of agency we should be interested in: actions that are done for reasons, and therefore subject to the sense of the question "Why?" that aims to elicit those reasons. On this approach, the nature of intentional action is inseparable from the role it plays in the practice of asking for and offering this kind of explanation. And even those who would not limit the scope of action theory in this way consider rational agency to be a crucial aspect of our topic. Focusing our attention on the kind of thing that we explain by appeal to reasons will therefore be a good way to get the subject matter into view. So: how exactly do the reasons cited in answer to the question 'Why?' explain what an agent is doing or has done?

1. Guises of rationalizing explanation

We can begin by noticing that rationalizing explanations of action can take several different forms. Much of the philosophical disagreement in this area concerns which of these forms, if any, is the true or most basic one. First, some rationalizing explanations explicitly cite a consideration that was the agent's reason for acting:

> *Reasons*: "Mariko's *reason for* eating a banana peel is/was that eating fibrous things is good for the digestion.

Reason-citing explanations pick out some feature of the action that justifies it or otherwise renders it intelligible, at least from the agent's perspective. This is often what we are looking for when we ask why a person did something – we want to know why they thought their action was choiceworthy or made sense to do.

Not all rationalizing explanations are explicitly framed in terms of the reason for which the agent acted, however. A second way in which we can rationally explain an action is by showing how it is related to some further goal or end the agent has:

> *Teleological*: "Mariko ate/is eating a banana peel *in order to* get more fiber in her diet.

Teleological explanations are characteristically neutral with respect to the ways, if any, in which the agent viewed her action as good or desirable. Rather, they elucidate the action in purely means-end terms, revealing what the agent ultimately aimed to accomplish in undertaking that action.

Third, we might explain the action by pointing to certain psychological features of the agent, thereby revealing her understanding of the world and the motivations behind her deed:

> *Psychologistic*: "Mariko ate/is eating a banana peel smoothie because she *wants* to lose weight and *believes* that eating less fruit and more fiber will cause her to lose weight."

The psychologistic form of explanation highlights the importance of what the agent had in mind. After all, a reason cannot explain an action if the agent did not believe it or was not motivated by it. It might be that blending a banana peel is the best way to sharpen the blades of one's blender, but if Mariko did not know this, or did not care about blender maintenance, then the blade-sharpening properties of her action do not explain it.

Fourth, we might simply cite another action that the agent was engaged in:

> *Naïve*: "Mariko walked/is walking into the kitchen because she *is getting* a snack."

We can call this form of action explanation "naïve" to contrast it with the psychologistic form, which we might view as more "sophisticated." Naïve action explanation recharacterizes the action to be explained in terms of another action that is or was in progress. It reveals that on this occasion, Mariko's act of walking into the kitchen is also an act of getting a snack. All four examples are

couched in both the past tense and the present progressive, in case it ends up mattering whether the action is still ongoing when the explanation is offered.

Though distinct from one another, each of the guises emphasizes an aspect of the rational intelligibility of action that tends to be implied by the others. The Teleological guise brings out the way in which rational actions tend to be goal-directed, undertaken as a means to some further end. We gain understanding of what the agent is doing when we discover what that end is. The Psychologistic guise brings out the fact that the end must be something that the agent herself has in mind and is moved by, as well as that the means-end relationship might only exist in her mind (if she is mistaken about whether her action will help her achieve her goal). We gain understanding of her action by seeing it from her perspective. The Reasons guise brings out the fact that the agent will normally be motivated to pursue an end because she sees it as desirable, good, or just something that it makes sense to do. It increases our understanding by revealing why the agent has the end that she does. And the Naïve guise invites us to take a step back and view what is going on as a smaller part of a larger whole.

2. Reasons for action: motivating vs. normative

Talk of the "the reason for which the agent acted" needs to be clarified in a few different ways. First, this is a philosophical term of art, and is meant to be compatible with the thought that we often act for multiple reasons. It is possible to donate to charity both for the reason that people are in need of help and for the reason that it will reduce your taxes.

Second, the phrase is meant to get at what some philosophers call "motivating reasons:" the reasons that actually move us to act. This is contrasted with "normative" or "justifying" reasons: considerations that really are good reasons for action. This distinction is drawn because normative reasons can sometimes fail to motivate us, while motivating reasons can sometimes fail to be good reasons. In the above example, we can imagine that Mariko has a good normative reason to eat the banana because bananas are a source of needed potassium. She is not motivated by this consideration, however, either because she does not know that bananas are a good source of potassium or because she cares more about avoiding calories than getting the nutrients she needs. On the other hand, her motivating reason for eating the peel might not be a good normative reason, if she does not in fact need more fiber in her diet.

Calling some reasons motivating and others normative can be misleading, since it sounds as though there must be two different categories of reasons. And indeed, some philosophers think that there are. But this may not be the best way to conceive of the terrain here. After all, we should try to avoid the rather depressing conclusion that we are *never* motivated by good normative reasons. We can do this by holding that motivating reasons and normative reasons are the same kind of thing – that they are both 'considerations', say. When we call a consideration a motivating reason, we are focusing our attention on what mattered from the agent's perspective, whereas when we call it a normative reason, we are focusing our attention on whether the agent was right to view things in this way. At any rate, the philosophy of action is primarily concerned with reasons in their motivating guise.

That said, there is philosophical disagreement over whether there are normative constraints on what kinds of things can feature as motivating reasons and thereby serve as rationalizing explanations of action. Can just *any* consideration be the reason for which the agent acted, regardless of how bizarre it is? Or must motivating reasons bear sufficient resemblance to good normative reasons if they are to explain action? To illustrate the question, consider an example of Anscombe's in which a person claims to want a saucer of mud – not because he wants to sculpt with it, or use it in a facial, or plant a garden, but simply for its own sake. She asserts that we can make no sense of this professed "desire," and to that extent, we might think that no mud-seeking action can be explained by it. Other illustrations involve taking highly peculiar means to one's ends. If a person is motivated by thirst to put a dime in a pencil-sharpener, or chooses to have coffee out of love for Sophocles, we might think that he fails altogether to count as acting for a reason. There are those who deny, however, that there are any such inherent constraints on what we can desire or find motivating about an action. As long as the consideration did in fact motivate the agent, no matter how bizarre, we can cite it as his reason in explaining why he acted as he did.

3. More on the 'Why?' question

If we are using rationalizing explanations as a pre-theoretical tool to understand human agency, we should aim to be clear about exactly what the 'Why?' question is asking. Outside the context of intentional action, the explanations we seek are often answers to the question 'Why did some event X happen?' For instance, we might ask 'Why did the plane crash?', seeking an answer like 'Because the angle-of-attack sensors malfunctioned.' This kind of explanation tends to cite a cause or condition that contributed to the fact that the event occurred, as well as to the fact that it occurred when it did. Roughly, the implication is that if the sensors had not malfunctioned when they did, the plane would not have crashed on that particular occasion. Among other things, this kind of understanding of past events helps us to predict what will happen in the future, such as that there will be more crashes involving planes of this type.

As we have already noted, when we inquire into why an agent acted as she did, we seek illumination of what it was that made the action seem to the agent to be sensible or worth doing. But in addition, it is plausible to suppose that we are also asking why the action happened at all. We want to know not only what Mariko was trying to accomplish by putting the peel in the blender, but also why she was making a smoothie at that time rather than studying or doing nothing. Thinking of the explanandum – the thing to be explained – in this way has the appeal of locating human actions in a broadly naturalistic order, in which they are on the same footing as other events that we aspire to explain and predict. If action explanations can tell us something about why a particular action occurred, then they can also help us to predict what people will do in the future.

Some philosophers deny, however, that explaining why something happened is part of what action explanations aim to do. For instance, Jennifer Hornsby writes:

> For even where there is an event of the agent's doing something, its occurrence is surely not what gets explained. An action-explanation tells one about the agent: one learns

> something about her that makes it understandable that she should have done what she did. We don't want to know (for example) why there was an event of X's offering aspirins to Y, nor why there was the actual event of X's offering aspirins to Y that there was. What we want to know is why X did the thing she did—offer aspirins to Y, or whatever. When we are told that she did it because she wanted to help in relieving Y's headache, we learn what we wanted to know (2004, 8).

The distinction Hornsby is making between "the thing X did" and "the occurrence of X's doing the thing she did" is subtle, and I say a bit more about it in Chapter Four, section 1. It suffices here that she is denying that action explanations should illuminate the particular occurrence of an action as distinct from contextualizing the type of action it was. Now, if the 'we' of this passage is taken as a sociological claim about what all people actually want out of an action explanation, I think this is mistaken. Rightly or not, many of us do want the occurrence of the event of the agent's acting to be illuminated. In any case, one's stance on the question of what we *should* want out of an action explanation is likely to shape one's thinking about which of the theories on offer is most plausible, and so it is worth some explicit reflection.

4. Action explanation: four views

With these preliminaries in place, we can now canvas four specific views about how rationalizing explanations work to explain action. The main disagreement is over which of the four guises of action explanation is more fundamental, and how that fundamental form is related to the kind of causal explanation that is familiar from the natural sciences. Are rationalizing explanations broadly in the same business as the explanation of why the plane crashed? Or are they tracking a completely distinct, non-causal structure?

a. The Rational Interpretation View

According to the first view we will consider, inspired by Gilbert Ryle and Ludwig Wittgenstein, the fundamental form of action explanation is the one referred to in section 1 as *Reasons*. No further account in terms of some deeper phenomenon is required to understand how reasons explain what rational agents do. Rather, the thought is that rational activity is a pattern that is explanatory in its own right. How does this work? Well, rational creatures by definition tend to desire what they believe to be good, to hold beliefs that are supported by the information they have, and to act in pursuit of what they desire given what they believe. These observations are meant to be conceptual rather than empirical in nature – it is simply what we mean by 'rational'. When we ask why an agent is doing something, according to this view, we aim to interpret that particular action within this broader pattern and thereby reach a deeper understanding or "Verstehen." We do this by redescribing the action in a way that reveals the considerations in light of which it was intelligible from the agent's perspective.

Suppose we see Gilbert driving off in his car late at night and wonder why he would do such a thing. The first step in interpreting his action is to use the context, behavioral evidence, and Gilbert's own reports in order to redescribe the deed in terms of a more general purpose: he is going to the store. We then attempt to locate what kinds of motivations and beliefs could have featured as the premises in Gilbert's reasoning that concluded in going to the store. If we learn that he wanted some chocolate, we can infer that he believed he could buy some at the store. We gain understanding of what he is doing because this is just the kind of thing a rational agent does when he wants some chocolate, has a car available, and has nothing else better to do. His action is explained by showing that it has the structure that is characteristic of rational activity. Call this the "Rational Interpretation" view of how action explanations work.

Of course, none of us is perfectly rational. A good interpretation of a person's behavior will take into account that human beings sometimes want things we shouldn't want, given what's good for us. We believe things we shouldn't believe, given our evidence, and we do things that are not especially well-suited to achieve what we want given our beliefs. Still, many of these rational flaws follow predictable patterns. We tend to prioritize short-term pleasures like smoking and tanning over our long-term health, while our beliefs are shaped by wishful thinking, confirmation bias, and base-rate neglect. Thus, the understanding provided by a reasons explanation need not depend on ideal rationality; the pattern can be stretched to accommodate common foibles. What it cannot allow for is the extreme irrationality of seeking to acquire mud for no further purpose or putting a dime in the sharpener because one wants a soda. Proponents of the Rational Interpretation view will see this as the right result.

But what makes a candidate rational explanation either true or false? As long as we can make good rational sense of Gilbert by attributing a desire for chocolate and a belief that there is chocolate to be had at the store, can there be any further question about whether this explanation is the right one? Those who favor this approach take the kind of understanding we seek to be primarily first-personal, and so will prioritize the agent's own report of her reasons for action. Indeed, a central idea found in the work of both Wittgenstein and Anscombe is that the agent's answer is not really a report of some independent fact, but rather an expressive act that has the power to *make* the connection between reason and action and thereby determine the fact of the matter. In other words, the reasons for which we act are normally what we sincerely assert they are. This first-personal authority is not absolute; it can be defeated by strong behavioral evidence that conflicts with the agent's own pronouncement. If we know that Gilbert doesn't have a sweet tooth and that he often finds excuses to visit the store when a certain cashier is on duty, we might well disregard his claim that he is motivated by chocolate. This kind of case must be the exception, however, in the sense that for Gilbert to be a rational agent at all, the answers he offers to the 'Why?' question must normally be accepted as decisive.

Most importantly, the Rational Interpretation view denies that rationalizing explanations have anything to do with identifying the *causes* of the agent's behavior. In this, it is motivated by the desire to avoid "scientism:" an excessive, even slavish tendency to think that science is the only genuine source of knowledge about reality. Proponents of the view have argued that because the rational patterns picked out by action explanations are logical in nature, the agent's desires and

beliefs could not be connected with her actions as cause to effect. Roughly, the thought is that 'action', 'belief', and 'desire' are conceptually interdefined: desires just are states that produce action, given our beliefs; beliefs just are states that guide action, given what we desire; and action just is the product of wanting something and believing we can get it. David Hume influentially claimed, however, that cause and effect must be "distinct existences:" among other things, effects cannot be logically inferred from their causes. Otherwise we would be able to know about causal relationships in the world without any kind of empirical investigation, and this is not in fact possible. Advocates of the "logical connection" argument conclude that entities standing in logical relations to one another cannot also stand in causal relations, and that beliefs, desires, and intentions therefore could not be causes of action. As A.I. Melden puts the claim, "… where we are concerned with explanations of human action, there causal factors and causal laws in the sense in which, for example, these terms are employed in the biological sciences are wholly irrelevant to the understanding we seek" (1961, 184).

To be clear, to be an "anti-Causalist" about action explanation need not involve thinking that the mind makes no causal difference to the world, or that mental states like belief and desire are epiphenomenal. The anti-causalist will generally be fine with claims like "Mariko's desire for a banana caused her to drool," in which what is caused by the desire is a mere physiological response. What she denies is that desires serve as causes of action when they are cited as a way of p[illegible]g out the agent's reasons. "We explain nature," wrote Wilhem Dilthey, "but we understand the life of the soul" (1961, 144).

b. The Causal Theory of Action Explanation

As the quotations from Melden and Dilthey illustrate, the Rational Interpretation model is discontinuous with the kinds of explanations we seek in the sciences. This does not necessarily mean that it is incompatible with them, but it does raise difficult questions about how the very same occurrence can be given more than one complete explanation. After all, actions do seem to be the kind of thing that we can study using the frameworks of biology, psychology, and neuroscience. If we can give a complete explanation of why some behavior occurred from a s[illegible]fic point of view, this threatens to render the contribution of a rational interpretation mysterious, if not entirely superfluous.

The Causal Theory of Action Explanation aspires to offer an account that is potentially more unified with the empirical sciences. The view is standardly credited to Davidson and his seminal 1963 paper "Actions, Reasons, Causes," although it actually has a long history that traces back at least to Aristotle. To clear space for the Causal Theory, let us take a closer look at the "logical connection" argument that led the original Rational Interpretationists to rule out a role for causes. Again, the thought is that because the concepts 'desire', 'belief', and 'action' are logically related to one another, desires and beliefs could not be causes of action, for causes cannot logically entail their effects. Davidson points out that this argument is actually invalid. The mistake arises from confusing facts about concepts (the ways in which we human beings think about the world) with facts about ontology (the way the world *is* independently of how we conceptualize it).

On a widespread way of thinking about it, causation is a *metaphysical* relation that holds between particular events no matter how we describe them, whereas logical relations hold only between concepts or descriptions.

For example, suppose that the loud thunderclap caused my cat Blinky to startle. This will remain true even if we change Blinky's name to Ringo, or if we describe the events as 'a sudden noise' and 'the black cat startled' – the descriptions do not matter, as long as they succeed in picking out the same events. We could even choose descriptions that are logically related to one another; for instance, we could refer to the thunderclap as 'the cause of Blinky startling' and assert the trivial logical truth that 'The cause of Blinky startling caused Blinky to startle.' But the fact that the two events are logically related under these particular descriptions does not show that they are not also causally related qua events. Thus, the fact that action explanations serve to render actions intelligible in a way that exploits certain conceptual interrelations does not rule out the possibility that there is also a causal relationship at work between the events we pick out by using these concepts in an explanatory context.

Having cleared the way for the view, we can now be more precise about what the proposal is. According to the Causal Theory of Action Explanation, the most fundamental form of rationalizing explanation is *Psychologistic*. The basic idea is that actions are explained in part by citing the psychological activity that caused them to occur. According to Davidson's version of the view, the relevant mental causes are a combination of something the agent believes and something she desires, or otherwise has a "pro-attitude" about. "Pro-attitude" is a term of art Davidson uses to refer to a very wide variety of ways in which one can be attracted to a state of affairs: one can desire it, judge it to be good, right, or beautiful, simply feel a yen toward it, and so on. In our earlier example, Gilbert has a pro-attitude toward eating chocolate and believes that he can acquire some at the store. Roughly, the claim is that together, these two mental states (sometimes called a "belief-desire pair") play a causal role in bringing it about that he drives to the store. Other versions of the causal theory may include additional mental states like intention as potential causal antecedents of action.

It is crucial to emphasize that in saying that psychologistic explanations of action are a species of causal explanation, we are not denying that they are also rationalizing. To succeed, the explanation must also justify what was done, or at least render it somewhat rationally intelligible. In this respect, the view aims to incorporate the insights of the Rational Interpretation model while at the same time going beyond it.

Davidson's reason for thinking that rationalizing explanations are a species of causal explanation is less an argument than a challenge to opponents: what else could they be? Rationalizing explanations do not just mention *some* consideration that favored performing the action; they provide the reason *for which* the agent acted. To see the difference, think of a case in which there are multiple reasons in light of which the agent could have acted, but where he in fact acted for only one of those reasons. Suppose Kunal's great-aunt has left him her fortune in her will and has made clear to him that she does not wish the doctors to take extraordinary measures to prolong her life. When she falls ill, Kunal has at least two reasons to prevent the doctors from resuscitating her: it will hasten his acquisition of her fortune, and it will respect her wishes. But it

might be that he does it only out of a desire for her fortune, even though he also desires to respect her wishes.

Davidson's challenge is to account for this difference between having a reason and acting for that reason.

> *Davidson's Challenge*: a successful theory of action explanation must explain what is added to the fact that S had reason R to perform action A when we say "S A-ed because R."

The Rational Interpretation view has difficulty meeting this challenge. The only constraint on a good rational interpretation is that it fits with the agent's circumstances, behavior, and her own assertions about what her reasons are. This poses a dilemma whenever there are multiple reasons on which the agent could have acted. Suppose we defer absolutely to the agent's assertion: if Kunal sincerely asserts that he acted out of respect, then it must be true. This would imply, implausibly, that we cannot be wrong about our own reasons. Plenty of Freud-inspired psychological research suggests otherwise; people frequently confabulate about their reasons for action without realizing we are doing so. Indeed, this is what the term 'rationalization' is ordinarily used to mean.

Alternatively, if we do not simply take the agent's word for it, there may well be no fact of the matter about the reason for which he really acted. True, the data available for interpretation may include the whole history of the episode, minor tics in his behavior, and so forth. But this does not ensure that there will be only one way to rationalize Kunal's deed, and insofar as there are multiple interpretations that fit equally well, they are all equally good as interpretations. Nothing further can be said about whether Kunal really acted because he wanted the money or because he respected his great-aunt's wishes, since attributing either of these reasons would make good rational sense.

Now, the proponent of the Rational Interpretation view might simply accept this result; if reasons-explanation is a primitive phenomenon, then there might be cases that bottom out in this kind of indeterminacy. This would be to reject the idea that there will always be a distinction between merely having a reason to act and genuinely acting on that reason. In contrast, the Causal Theory aims to vindicate this distinction by locating the difference in the causal history of the action. A reason the agent had, but didn't act for, is a belief-desire pair that did not in fact play a causal role in bringing her action about. And if she acted for a reason, then there must have been a relevant belief-desire pair that did play a causal role. If we seek to understand the real reason why Kunal prevented his great-aunt from being resuscitated, we must inquire into what actually caused his behavior. Proponents of the Causal view tend to see Davidson's Challenge as a requirement that must be met, and are skeptical that this can be done by an entirely non-causal form of explanation.

We have seen what the Causal Theory of Action Explanation is, in its broadest outline: that explanations of action mentioning the agent's beliefs and desires are not only rationalizing, but also a species of causal explanation. Let's now clarify what it isn't. Unfortunately, it is a widely

misunderstood view, both because the proponents of the view have not always been clear and because its opponents have tended at times to target a straw man.

i. The Causal Theory of Action Explanation is a partial answer to the question "How do reasons explain action?" It is not necessarily even a partial answer to the question "What makes an action voluntary or intentional?" For all that has been said, it is an open question whether being explicable in terms of reasons is essential to being an intentional action. We might think it is not essential, since it seems possible to act intentionally but for no particular reason – say, to intentionally avoid stepping on the cracks in the sidewalk (see section 5 below).

That said, most proponents of this view do think that having the right causal history is a necessary condition of being an intentional action. It does not follow that this is also a sufficient condition, however; it might well be that there are other essential components. Confusingly, this means that one can subscribe to the Causal Theory of Action Explanation without subscribing to a view that is sometimes called the Causal Theory of Action, namely, that being caused in the right way is both necessary and sufficient for being an intentional action. We will examine this ambitiously reductive view about the metaphysics of action in Chapter Four. The two views often get blurred together under the label "causalist," so it is important to note that they are in fact distinct.

ii. There are at least two different kinds of causation that we might be referring to in the context of agency. On one hand, there is what Aristotle called "efficient causation:" the primary source of change or rest. This is the kind of causation at issue in claims like "the cue ball caused the eight ball to roll into the pocket" and "the assassination of Archduke Ferdinand caused World War I to break out." More broadly, it is the kind of causation that the natural sciences take themselves to be investigating. On the standard way of thinking about efficient causation, the causal relation holds between two events: one event that is the cause, and a second event that is the effect. On this view, it is not strictly speaking the cue ball that causes anything, but the *event* of the cue ball striking the eight ball with a certain momentum. There are also variations of the view that accord a less fundamental role to events, allowing the causal relation to hold between facts or other features of a situation.

On the other hand, it has sometimes been argued that there is a distinct kind of "agent-causation" that is at issue when someone acts. The agent-causal relation does not hold between two events or facts; rather, the cause is held to be the agent herself, and not merely some event that she participates in or some proper part of her. In other words, if the cue ball were an agent, then it would be accurate to say that the cue ball itself was the fundamental cause of the eight ball's rolling into the pocket – not the movement of the cue ball, or the event of its striking the eight ball, but simply the ball, viewed as a substance. This is not an understanding of causation that is familiar from science, but the intuition is that an action must be caused by nothing less than the agent as a whole (stemming, often, from a conception of action that prioritizes attributability). We will briefly return to the Agent-Causal View in Chapter Four, but to be clear, the notion of causation at issue in discussing Davidson's view and its descendants is efficient causation and not agent-causation.

iii. The Causal Theory of Action Explanation is sometimes expressed using the slogan "reasons are causes." This is a potentially misleading phrase. The view holds that the causes of action are mental states like beliefs and desires – or more precisely, mental events that are the manifestations of such states. If "reasons are causes," then reasons for action would turn out in every case to be belief-desire pairs. But this is a highly unappealing implication, whether we mean "normative reasons" or "motivating reasons." If what we mean is "*normative* reasons are causes," then we are committed to thinking that the only considerations that count in favor of doing anything are facts about our own mental states. Read strictly, this is an unappealing view. As Jonathan Dancy points out, there are rare cases in which facts about our own beliefs and desires do provide normative reasons for action, as when the fact that you believe the CIA is spying on you, or the fact that you desire to hide under your covers all day, are reasons to consult a psychiatrist. These cases are very different, however, from the normal case in which our reasons are non-mental facts. It is the fact that fiber is good for the digestion, together with the fact that banana peels contain fiber, that gives Mariko a reason to eat a banana peel.

It might seem that the view does better if we understand the claim as "*motivating* reasons are causes." After all, the main impetus for introducing the category of motivating reasons is to capture the agent's perspective on her action, whether or not that perspective is mistaken. Further, the notion of "motivation" and the notion of "cause" seem to be connected. Why not then say that all motivating reasons are mental states, whereas normative reasons are worldly facts? But as noted in section 2, dividing these two categories so sharply would make it impossible, or at least rare, to act for good normative reasons. The considerations that justify our actions would almost never be the reasons that actually motivate us. This is also an unhappy result.

One plausible solution is to hold that motivating reasons are the contents of our attitudes: not our beliefs and desires, but *what* we believe, in combination with *what* we desire. What rationalizes Mariko's smoothie-drinking is the fiber in the peel, as represented by her beliefs, and the advantages of good digestion, as represented by her desires. When our beliefs are true, and when we desire what is genuinely desirable, then our motivating reasons are also good normative reasons. And when our attitudes are out of kilter with reality, our motivating reasons are not good normative reasons. Now, it is an open question whether and how the contents of our attitudes themselves figure in the causal explanation of action in addition to the attitudes themselves. But we should not be confused into thinking that according to the Causalist, the only reasons we ever act on concern our own beliefs and desires.

iv. A final complication concerns the topic of mental causation more generally. Is it plausible to suppose that mental states like beliefs and desires are ever the true causes of physical events like bodily movements? Assuming that (in human beings, at least) the mind is realized by the brain, the particular states that we pick out using mental concepts like 'belief' and 'desire' are neural states. It is thus tempting to think that the causal story of how our bodily movements are initiated and guided must ultimately be couched in terms of neural events and other physical states of the organism, rather than in the vocabulary of "folk psychology." It is still very much an ongoing

debate in the philosophy of mind whether or not familiar mental concepts or properties will feature in a mature account of the causes of human behavior.

It may seem that the Causal Theory of Action Explanation prejudges this question by insisting that beliefs, desires, and other pro-attitudes are causes. This is not necessarily so. Davidson himself denies that mental concepts figure in causal laws, and takes mention of mental states like belief and desire to be mere shorthand for some associated event that was the catalyst of the relevant behavior. The Causal Theorist is committed only to thinking that when we give rationalizing explanations of action, there is a causal connection in the vicinity that partially grounds the truth of the explanation, even if that connection can only be formulated precisely using non-mental concepts. In this respect, it is not necessarily a reductive view about the relationship between the mental and the physical (though any given version might well be).

c. Teleological Realism

Davidson's Challenge is not really an argument in favor of the Causal Theory. It simply asserts that the Causal Theory is the only game in town and places the burden on the opponent to prove otherwise. Assuming that the Rational Interpretation view does not meet the challenge, are there any other games in town?

The Causal Theory can look inevitable if we take the Psychologistic form of explanation to be primary. An alternative approach rejects this first step, and instead takes *Teleological* to be the basic one. Teleological explanations, recall, have the form 'S A-ed in order to B' or 'S's A-ing was directed at B'. According to the Causal Theory, when claims like this are true, it is because of some underlying fact about the agent's psychology – that S wanted to bring about B, or believed that A-ing was a way of bringing about B. In contrast, the view we will call Teleological Realism denies that such teleological facts are reducible to some other, non-teleological phenomenon. Rather, it accepts that teleology is part of the basic furniture of the world.

The Teleological Realist view has been most prominently defended by George Wilson and Scott Sehon. The basic idea is that intentional action is behavior that is "sentiently directed" at some goal, and that the 'because' of action explanation functions to reveal what that goal was. When we ask why Kunal prevented the doctors from resuscitating his great-aunt, we are aiming to elicit the further end at which his action was directed. The most straightforward way of answering the question will be to say that he did it *in order to* inherit the money, but the other guises of explanation can be understood as oblique ways of specifying the relevant goal. Speaking of what Kunal desired and believed, for instance, is simply a different way of illuminating the state of affairs his behavior was directed at achieving. Importantly, though, "directing at" is not meant to be an efficient-causal notion; the answer to the question of what Kunal's behavior was directed at is not also an explanation of what caused his mouth to open and utter the words "do not resuscitate." The latter question will, presumably, be answered in neurophysiological terms.

Skepticism about the scientific validity of folk psychology, which understands the mind in terms of attitudes like desire, belief, and intention, is a major motivation for this approach. If a completed science of the mind and related phenomena such as motor control will not involve

anything recognizable as beliefs and desires, then we might well think that any causal explanations invoking such items will turn out to be mistaken (though Davidson famously disagreed with this inference; see his (1970)). Rather than concluding that our common-sense action explanations are completely off the mark, however, the Teleological Realist proposes that we understand them in non-causal terms. If their explanatory power derives from teleology rather than causation, then they need not be in competition with scientific explanations.

A second motivation for the view stems from skepticism that causation could provide the illumination that we seek from action explanations. Teleological Realists like Wilson and Sehon have charged the Causal Theory with suffering from a fatal flaw known as the problem of causal deviance. In essence, the worry is that being caused and rationalized by a belief-desire pair is not enough for an event to have the status of an intentional action, if the event is caused *in the wrong way*. In Davidson's own classic example, a mountain climber has lost his hold and a second climber is holding him on a rope, at great peril to his own life:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally (1971, 79).

The point is that the causal chain leading from his belief and desire to his letting go of the rope is "deviant," relative to the way the Causal Theorist thinks actions are normally caused by our beliefs and desires. This is not merely a problem for causation by belief and desire. Although Davidson's example makes it sound as though the problem is the absence of a choice, we could easily construct a similar example in which the climber does make a choice, or form an intention, and it is the choice or intention that so unnerves him.

What exactly is this kind of case supposed to show? Things get tricky here. Most directly, the problem of causal deviance is a problem for a different kind of view – a causal theory that attempts to provide sufficient conditions for being an intentional action. If the idea is that *what it is* to act intentionally is for one's bodily movements to be caused in the right way, it is essential to make good on the clause "caused in the right way." Any case in which the proposed sufficient conditions for being an intentional action are met, but where the resulting behavior fails to be intentional, is a counterexample to the proposal. We will discuss this challenge in more detail in Chapter Four, when we discuss a theory of this variety.

However, the topic here is the Causal Theory of Action Explanation and not a causal analysis of what actions are. It is not as obvious how deviant causal chains are supposed to demonstrate the hopelessness of the former view. On the face of it, that view is committed only to the following claim: *if* an event is the kind of thing that can properly be given a rationalizing explanation, then the truth of the explanation depends partly on the right kind of causal connection holding between the agent's attitudes and her behavior. In cases of causal deviance, the Causal Theorist may simply say that the resulting event is not subject to a rationalizing explanation – after

all, the example stipulates that it happened because the climber was nervous and lost control. The view is simply not committed to counting everything that is both caused and rationalized by the agent's beliefs and desires as an intentional action. Thus, the problem is not that these cases are counterexamples to the theory.

Still, the possibility of causal deviance does put pressure on the idea that causation plays any significant role in rationalizing explanation. It turns out that causation alone is not enough to meet Davidson's Challenge of explaining the difference between having a reason and acting for that reason. The possibility of causal deviance shows that in addition, the causal connection must not be of the wrong kind. And one worry is that by "right kind of cause," we simply mean "whatever it takes to act for a reason," in which case the appeal to causation isn't helping to explain anything. Further, if the Causal Theorist makes the move just proposed, she will be granting that acting intentionally is not just a matter of being caused to act by the mental states that represent one's reasons. This may lead us to worry that the other non-causal features of intentional action, whatever they are, are really doing the explanatory work.

Going in for Teleological Realism neatly sidesteps the problem of causal deviance, since the view has no particular causal commitments. However, it does this at the cost of the more straightforward naturalism of the causal view. We now need an answer to the question "What makes teleological explanations true?" Now, some teleological facts are perfectly explicable in naturalistic terms because they are grounded in non-teleological facts. For example, the function of the heart is to circulate blood, or as we might say, its pumping is directed at the goal of circulating blood. But there is hope, at least, of explaining how this could be true in virtue of the process of natural selection; it is not simply a brute fact. Likewise, your alarm might sound in order to wake you up at 7am, but this is true in virtue of a person having programmed it with this intention. In contrast, according to the Teleological Realist, the fact that Bas went to the kitchen in order to get some tea is irreducible: it is not grounded in, or explicable in terms of, any non-teleological fact. As Sehon writes, "When making evolutionary explanations, any appeal to [teleology] is merely heuristic and can be spelled out in purely casual terms. Teleological realism claims that no such reductive story can be told concerning the typical teleological explanations of CSP[common-sense psychology]" (2005, 153). This claim seems to require that we add teleology to the basic furniture of the world, not because science demands it but because our common-sense action-explanations do. This is the aspect of the view that conflicts with at least some forms of naturalism.

Further, irreducibly teleological explanations are mysterious because they seem to appeal to a future state of affairs to explain something that is happening now. 'Having tea' is the state of affairs that is meant to explain Bas's going into the kitchen, even though Bas does not yet have tea. Stranger yet, he might never have tea, since he might find once he gets to the kitchen that he is out of tea. This is strange because we usually think of explanations as *factive*: P cannot explain Q unless P is actually the case. If we find out that Ivy doesn't have measles, then measles cannot be the explanation for her rash. And while according to the Causal Theory, it might well be that the agent's beliefs and desire are mistaken, she must actually have those attitudes if they are to explain what she did. But in the case of teleological explanations, the claim is that a state of affairs that is

not yet the case, and may never be the case, somehow explains the agent's current behavior. This is a significant mystery at the center of the teleological view.

A final question is whether this view has a convincing response available to Davidson's Challenge. Sehon argues that when there are multiple candidate goals that might explain an agent's behavior, we can establish which of them is the true explanation by examining a set of counterfactual circumstances that tease the different goals apart. To return to the example of Kunal, we can ask what Kunal would have done if honoring his great-aunt's wishes were not also a way of hastening his inheritance, or if hastening the inheritance had required that he violate her requests. Now, it is surely correct to point out that certain counterfactual patterns are implied by the claim that an agent acted for one reason and not another. The question is whether this pattern is in itself explanatory in the way needed to meet Davidson's Challenge. If the challenge were to answer the epistemological question of how to go about establishing which explanation is true, the appeal to counterfactuals seems apt. But from a metaphysical perspective, where the challenge is to explain what makes a reason-explanation true, we might think that the counterfactuals themselves do not ground the difference between having and acting for a reason. Rather, they are a symptom of whatever does ground that difference. The Causalist will agree that these patterns hold, for instance, but claim that they hold in virtue of certain of the agent's mental properties (or physical properties in the vicinity) playing the relevant causal roles.

d. Naïve Action Theory

In some ways an updated version of Teleological Realism, Naïve Action Theory (most prominently defended by Michael Thompson) holds that the most fundamental form of action-explanation explains one action with another: "I am doing A because I am doing B." For example, instead of saying that Bas is going to the kitchen because he *wanted* some tea and *believed* he could get some in the kitchen, or that he went to the kitchen *in order to* get some tea, we can simply say "Bas is going to the kitchen because he *is getting* some tea." Whereas the Teleological Realists are primarily motivated by the problem of causal deviance, the Naïve Action Theorist has a different objection that applies equally to the Causal Theory and the original formulation of Teleological Realism. Naïve action explanations, in which one action is explained by appeal to another, stand opposed to what Thompson calls "sophisticated" explanations that cite desires, beliefs, and intentions. These are often called 'propositional attitudes', since they are thought to take propositions as their contents: Bas believes that {there is tea in the kitchen}, and desires that {he has a cup of tea}.

The problem is that when it comes to action, it is a strain to interpret what is desired or intended in terms of some proposition. Rather, the object of desire or intention in these cases is normally expressed as an infinitive. Bas intends *to walk* into the kitchen and to get some tea, not *that* he walks into the kitchen or *that* he has some tea. The intuition is that Bas will not have done what he intended if he simply ends up in a situation that can be described as "Bas has tea" through no activity of his own. If he absent-mindedly lets his mouth hang open and a friend impishly pours tea into it, this should not count as Bas's action of getting tea. The general point is that propositions

represent states of affairs, but actions are not states; insofar as they have not been completed, they are things that are *in progress*. At the very least, this challenge shows that it will not be a simple matter for proponents of "sophisticated" psychological explanations to understand the kind of representation at issue.

Thompson draws a stronger conclusion, arguing that intending is not properly understood as a mental state or attitude at all. Rather, to intend an action is already to be in the process of doing it. The type of explanation of action at stake in action theory, as he puts it, "is uniformly a matter of locating the action explained in what might be called a developing process; it is just that this progress, development, or 'imperfection' must be understood to exhibit various types or grades" (2008, 132) To speak of something the agent *is doing* – 'Larry is hammering the nail because he is building a tree house' – is to assert that the action of building a tree house is well on its way. If the tree-house construction has not progressed much beyond its conception, we are more inclined to speaking of *wanting* or *intending* to build a tree house. But according to Naïve Action Theory, these are not references to psychological causes; rather, they are simply a way of expressing that the action is still in its infancy. It would be true, if linguistically awkward, to use a naïve formulation even in such cases: 'Larry is buying lumber because he is building a tree house this weekend." As Thompson writes, using his own example of buying eggs to make an omelet, "The unity that joins egg-purchase to omelet-making, thus narrowly construed, is the unity that joins the acts we are willing to call parts of omelet-making to one another, and makes an intentional action out of them" (2008, 132) Naïve explanations function to reveal this unity by relating one action – the buying of lumber – to a more encompassing action – the building of a tree house – of which the lumber-buying is a part.

In directing our focus to the distinctive features of actions in progress, and the logic of the imperfective language we use to describe them, Naïve Action Theory has made a profound contribution to contemporary philosophy of action. But although it offers an account that is in some ways more nuanced than the original Teleological Realists, it faces many of the same worries. Where the Teleological Realist speaks of goals or functions explaining the processes by which they are achieved, the Naïve Action Theorist speaks of the form of the act-process explaining its own parts. It is because Larry is building a tree house, and because hammering is part of tree-house-making, that he is hammering the nail. But this idea is at least as puzzling as the purely teleological version, and will similarly force us to say in some cases that the parts of the action are explained by a whole that never actually occurred. For instance, if Larry never gets around to building the tree house, his buying of lumber must be explained by a tree-house-building that never happened. Perhaps this is not so odd, since there is a sense in which on this view, the buying of the lumber *just is* a tree-house-building that is underway, though still in its very early stages. It is more difficult to accept, however, in cases where there was never any possibility of succeeding. If Jane Eyre is donning a wedding dress with the idea of legally marrying Mr. Rochester, who is already legally married to someone else, there seems to be no sense in which she is already marrying him in dressing herself. This would be, absurdly, to credit her with already doing something that it is not possible for her to do.

Finally, we should ask how the Naïve Action theorist might answer Davidson's Challenge. To formulate the challenge in naïve terms, the task is to explain the difference between performing some action X that is in fact a part of some other action Y that one is engaged in, and doing X *because* it is a part of Y-ing. Larry might currently be involved in several home improvement activities of which buying lumber is a part, and yet he might be buying lumber only because he is building a tree house. In virtue of what is this the case? Thompson seems to suggest that the answer lies in the agent's thought about the connection between X and Y. Larry is buying lumber because he is building a tree house if and only if he thinks of buying the lumber as a stage of building the tree house. If this is right, it does provide an answer to Davidson's Challenge. We might worry, however, that this answer comes at the cost of the ability to question the agent's own perspective on what she is doing and why. As I see it, the intuition behind the challenge is in part that we do not have perfect or even especially good access to the reasons for which we act, especially in cases where the protection of our ego is at stake in the answer. It is commonplace, for instance, to think of oneself as acting for a more virtuous reason than one really is. But this apparent truism must turn out to be false if an agent's reasons are simply whatever he thinks they are.

5. Arational Action

Stepping back from the question of how action explanations work, let us return briefly to the question of whether all intentional actions are amenable to some form of rationalizing explanation. It is tempting to suppose that the answer is 'Yes', since this would offer hope of understanding action entirely in terms of rational explanation. But there is good reason to think that there is such a thing as "arational action." For one thing, even though Anscombe emphasized the centrality of the 'Why?' question for understanding intentional action, she held that a legitimate answer to the question is "No reason." In other words, to claim that an action was done for no reason does not show that it was unintentional on her view. Rosalind Hursthouse offers numerous vivid examples of such cases, many of which are done out of emotion rather than out of some belief about the desirability or usefulness of the act. One might shout at or kick a footstool that was in one's way, or use a cigarette to burn out the eyes of an ex-lover in a photograph. It is implausible in these cases to attribute a belief to the agent that such acts make good sense, accomplish anything useful, or are in some other way fitting or morally good. Nor is it especially plausible that all such actions are motivated predominantly by the desire to express one's emotions. That said, they are not done unintentionally; the agent is sufficiently in control of herself and knows what she is doing.

One might protest that such examples show only that there must be no *further* reason for which the action was done, not that they are done for no reason. The reason is that the agent wanted to do it, just for the sake of it – not because it will be pleasurable, or cathartic, but "just because." The risk here is that we are merely saying something trivial, adding nothing of explanatory value. There is a thin sense of 'wanting' that is simply entailed by the fact that an action was intentional – if the agent A-ed intentionally, then she wanted to A. But it does not follow from this thin sense of wanting that the agent had any desire to A that was antecedent to

her doing it, or that she saw anything to be said in favor of A-ing. Thus, it is not clear that "because she wanted to" should itself count as a rationalizing explanation.

Summary

We started with the insight that intentional actions are subject to a distinctive kind of explanation. When we ask 'Why are you doing that?' or 'Why did you do that?', a positive answer usually takes one of four forms: it cites a reason for action, mentions something that the agent wanted or believed, makes reference to some further goal the agent had, or invokes something else that the agent is doing. The question is how these answers explain the action, and which (if any) of them is fundamental.

The Rational Interpretation view fits best with the conviction that action explanation does not aim to explain why some event occurred or what motivated the agent in a mechanistic sense. It holds that these explanations work by assimilating the action to the kind of pattern that rational creatures tend to exhibit: broadly, the pursuit of what is good, fitting, or sensible, given what they believe about the world. We find out why this person did what she did by coming to understand why any rational person could have viewed the action as worth doing. This kind of view will tend to think that it is constitutive of being an action that it can be made rationally intelligible.

Proponents of the Causal Theory agree that this is one aspect of action explanation but object that it is not enough. On this view, explanations of action do aim in part to explain the occurrence of the action qua event. Its central tenet is that while having beliefs and desires that rationalize doing A suffices for having a reason to A, one does not act for that reason unless the relevant beliefs and desires causally contribute to bringing it about that one A's. This view aspires to make action theory broadly continuous with naturalistic psychology.

The Teleological Realist takes the teleological guise of action-explanation to be fundamental. Action explanations work by revealing the ends to which the action in question is (putatively) a means. It is motivated by skepticism that invoking causation adds anything useful to action explanation, since the problem of deviant causal chains shows us that the requisite causal connection must be "of the right sort." Once we figure out the further conditions that rule out deviant cases – if this can even be done – the suspicion is that the causation part isn't doing any work. The Teleological Realist advocates for taking the phenomenon of sentient goal-directedness as basic and irreducible rather than trying to explicate it in causal-psychological terms.

Finally, Naïve Action Theory aims to understand actions as parts of other actions that are in progress. Not only does it hold that the fundamental guise of action explanation is the naïve form, but also that what we aim to explain are actions that are currently underway rather than deeds that have already been done. Like the Teleological Realist, it denies that the agent's mental states play a causal role in explaining action, but for the reason that intention and action are not static things; on this view, they have their own processual form that cannot be assimilated to the traditional model of propositional attitudes.

Suggested Reading

The first part of G.E.M. Anscombe's monograph *Intention* compellingly makes the case that being subject to reasons-explanations is a distinguishing characteristic of the kind of thing we seek to understand. The anti-causalist Rational Interpretation view was most ascendant in the mid-twentieth century. It was defended in a series titled *Studies in Philosophical Psychology*, often referred to as the "little red book" series. These include A.I. Melden's *Free Action*, Peter Winch's *The Idea of a Social Science and its Relation to Philosophy*, and Anthony Kenny's *Action, Emotion and Will*. Donald Davidson's famous rejoinder on behalf of Causalism can be found in "Actions, Reasons, Causes." For an overview of the disagreement over reasons and causes, see Giuseppina D'Oro and Constantine Sandis, "From Anti-Causalism to Causalism and Back: A Century of the Reasons/Causes Debate."

The locus classicus of the revival of the Causal Theory of Action Explanation is Davidson's "Actions, Reasons, Causes." His paper "Freedom to Act" is also relevant and contains the original discussion of deviant causal chains. For defenses of Teleological Realism, see George Wilson's *The Intentionality of Human Action* and Scott Sehon's *Teleological Realism*. And for further exploration of Naïve Action Theory, see Michael Thompson's *Life and Action*.

Jennifer Hornsby's "Agency and Actions" articulates general skepticism about the idea that action explanations are explanations of why some event occurred. Chapter One of Jonathan Dancy's *Practical Reality* offers a useful discussion of the distinction between motivating and normative reasons. And for more on what reasons are and how they are related to action, Eric Wiland's book *Reasons* is a great resource. Finally, Rosalind Hursthouse's "Arational Action" is an important reminder that not all intentional actions are obviously subject to traditional rationalizing explanations.

4. The Ontology of Action

What in the world *is* an intentional action? As we saw in Chapter Three, a first step to answering this question is to note that intentional actions are the kinds of behaviors that can be explained by citing reasons for action. This insight helps point us to our subject matter, but it does not tell us exactly what we are looking at. In this chapter, we will delve more deeply into metaphysical and ontological questions about the nature of action.

'Metaphysics' is the philosophical study of the fundamental nature of things, and 'ontology' is the branch of metaphysics concerned with the nature of being or existence. One major metaphysical question concerns how agency and action fit in with the rest of the natural world. Can we understand action using only those categories that we are independently committed to by the empirical sciences? Or must we introduce new resources, such as a distinctive form of causation, to account for the possibility of genuine agency? Views that aspire to the former approach are often called "reductive," while the latter approach is "non-reductive," though we must be careful here to distinguish between ontological reduction and conceptual reduction (see Chapter Two, section 12).

A second question concerns what kind of thing intentional actions are. We can ask this question at various levels of abstraction. Are we talking about events, or some other kind of logical category? And if they are events, what are they events of – willings, bodily movements, or some further miscellaneous collection? Third, and perhaps most importantly, we are seeking an answer to the question "In virtue of what are such things intentional actions?" In other words, once we have identified the kind of thing that can amount to an intentional action, we should like to know what grounds the fact that it is an intentional action. Is there some intrinsic property that is shared by all intentional actions and lacked by otherwise similar occurrences that are not instances of intentional action? Do they have a distinctive "form" or essence? Or does it depend on how a candidate event or occurrence was brought about? Finally, we will briefly discuss kinds of intentional action that may require a distinctive treatment in our theorizing: omissions and mental actions.

1. Which things in the world can be actions?

Let us start with the question of what actions are at the most abstract level. Though there is a healthy debate on this topic (as with most things in action theory), I will tend to refer to actions in what follows as a species of *event*. In doing so, I hope not to beg too many important questions, so I will explain briefly what I take to be at issue here.

As mentioned in Chapter Two, there are some philosophers whose approach to action theory is restricted to the first-personal "practical perspective" – the perspective of an agent engaged in deliberation about what to do. These philosophers reject the idea that actions are happenings that can be understood in more impersonal, empirical terms, and so would consider

most of the questions of this chapter to be fundamentally misguided. Setting these theorists aside, all others should agree that actions belong in the general category of things that happen. In this respect, they are unlike objects or states, which are, well, static: they exist rather than happen.

We can further distinguish between particular actions and types of things one can do. Particular actions are individuals rather than abstractions or universals – they happen once, at a time and a place (though they might be extended over large spans of time and space). I brushed my teeth today, in the bathroom, at 7:39am. My brushing my teeth today is not the same action as my brushing my teeth yesterday. On the other hand, we sometimes use the word 'action' to refer to thing that I did, namely, 'brush my teeth'. 'Brushing one's teeth' or 'going for a run' are kinds of things one can do; they are not particular events, but something like general blueprints for action. Thought of as a kind of thing one can do, the action of teeth-brushing I performed today *is* the same as the one I performed yesterday. Though the event of an agent's acting is also the doing of a kind of action, it is important not to run these two things together; we will get confused if we end up thinking that what agents do are events. Rather, agents do types of things, and the doing of those things on particular occasions are events of an agent's acting. And of course, any particular event can be the doing of several kinds of things at once, as when one brushes one's teeth while running.

Some philosophers contend that action theory should be primarily concerned with things done and not with the events of an agent's doing them. This move might help to avoid some of the puzzles that come up in this chapter. I will not restrict the discussion in this way, however, since many others do take these puzzles seriously. Thus, I will assume here that part of the task is to understand actions as, in part, spatiotemporal happenings (though not *mere* happenings). Beyond this, we can dispute what kinds of spatiotemporal happenings they are. Candidate categories, in addition to the general category of 'event', include processes, activities, continuants, and transactions. We will not go into great detail here about the differences between these proposals, since I tentatively think that the notion of event can be understood broadly enough to include most of them as sub-categories. Characterizing them as processes or activities emphasizes the fact that actions almost always take time to perform, and that they evolve from being incomplete to completion. They can be intermittent, as when one writes a book over the course of a year (or five) without writing at each moment. And they can change as they unfold; one can begin hiking cheerfully and enthusiastically and end up hiking dejectedly and dispiritedly. Events are not usually thought of as the kind of thing that can be incomplete, intermittent, or that can change over time, but it is open to us to use the term in an expansive way that allows for such qualities.

The most influential argument that actions must be a species of event is owed to Donald Davidson (though I should note that not everyone finds this line of reasoning persuasive). The argument is complicated and technical, but the basic idea is that thinking of actions as events solves a problem we would otherwise have with understanding the logic of how we talk about action (the "problem of variable polyadicity"). Suppose I tell you that "Jones buttered the toast at midnight with a knife." From this knowledge alone, you are licensed to make inferences to less committal claims like "Jones buttered the toast" and "Jones buttered the toast with a knife." The problem is to understand the logic in virtue of which these inferences are valid. Ordinary first-order predicate

logic does not get it right, since we can only infer from one sentence to another if they have a predicate in common. But in that system, predicates with different numbers of argument places must be different predicates. So if we interpret the sentence 'Jones buttered the toast' as predicating 'Buttered the toast' of 'Jones' – making 'Buttered the toast' a predicate with a single argument place – and the sentence 'Jones buttered the toast with a knife' as involving a two-place predicate holding of 'Jones' and 'a knife', then 'Buttered the toast' functions as two different predicates in these sentences. This means that we cannot validly infer the first sentence from the second.

Davidson's solution is to claim that action sentences like 'Jones buttered the toast with a knife at midnight' contain an implicit existential quantifier which quantifies over events. They assert that there was an event, and that this event had certain properties: it was an event of toast-buttering, done by Jones, done with a knife, and it took place at midnight. The list can go on indefinitely, since events can have any number of properties. More formally, the sentence has the following logical structure: ∃x (Buttered (Jones, toast, with a knife, at midnight, x)). Since all of t operties are predicated of the same event, the desired inferences go through straightforwardly an the problem of variable polyadicity is solved.

If all of that failed to make much sense, not to worry. The important point is simply that in order to make the logic of action sentences work, there must be a single metaphysical item that is the action and that can instantiate the various properties we want to attribute to actions – being done at a certain time, in a certain way, and so forth. Davidson claims that events are the kind of item in question, but as far as I can see, his solution does not actually depend on quantifying over events in particular. Insofar as processes and activities are not kinds of events, we could substitute either category in for 'x' in the above formula without impairing the overall proposal. In general, while the question of whether actions are events, processes, activities, or some other kind of thing remains a disputed issue, it seems to me that much of what is of interest in action theory is compatible with different views on this question.

2. Under a description

Assuming now for the sake of simplicity that actions are events, we can ask two further questions: which events are actions, and which actions are intentional? The beginning of an answer, I suggest, lies in how certain events are described. Events themselves do not come with any particular description attached; they simply happen. It is we human beings who pick out some events and describe them in various ways according to our interests. On at least one natural understanding of event individuation, we can refer to the very same event as 'the sacking of Istanbul', 'the sacking of Constantinople', or 'the end of the Fourth Crusade', depending on what we are interested in.

Whether an event was an intentional action depends on the description we use to pick it out. That is, the same event will be an intentional action under some descriptions but not under others. Imagine that Ivy is in the kitchen operating an electric mixer in a bowl of heavy cream. We may suppose that under the description 'making whipped cream' she is acting intentionally. But we can also describe what she is doing in other ways: making a terrible screeching noise with

the mixer, flicking bits of cream all over the walls, and so forth. Presumably, Ivy is not doing these things intentionally, even though her making whipped cream is the same event as her making a mess and a racket. In general, the slogan is that an event is an intentional action or not only "under a description."

What, then, is the relationship between actions and intentional actions? On one prominent way of thinking about it, an event is an action if there is at least one description under which it is intentional. So because Ivy is intentionally making whipped cream, this event is still an action of hers even if we describe it in ways that pick out unintentional features – the mess, the racket. After all, if her mother inquires who made the mess or the bothersome noise, the correct answer is 'Ivy did.' These things are attributable to her, and she may be held accountable for them. On the other hand, if there is no description under which an event is intentional – if, for instance, Ivy simply trips and falls face-first into the whipped cream – then it is not an action of hers. Rather, it is something that happened to her, as patient rather than agent.

3. Basic actions

We now need to know exactly which events are candidates for being intentional actions. A further point of widespread (though not unanimous) agreement is that we can usefully distinguish between 'basic' or 'primitive' actions and complex actions. The notion of a basic action has been cashed out in a variety of ways, but the best version of the idea is that basic actions are done "directly" or "immediately." They are not done by doing anything else. Complex actions, in contrast, are done by performing basic actions. 'Building a house' is an example of a complex type of action; it is not something limited beings like ourselves simply do directly or at will. Rather, it can be decomposed into a set of more primitive act-types such as framing walls, hammering nails, and laying shingles. These more primitive act-types can be broken down in turn into even simpler parts. But eventually, the thought goes, this regress must stop. There must be actions that can be accomplished without intentionally doing anything else that is undertaken as a means. Otherwise, an infinite number of actions would have to be performed in order to act at all.

Not everyone agrees that this conclusion is unacceptable. Michael Thompson has argued that there is no such thing as basic action precisely on the grounds that each intentional action can in fact be divided into infinitely many parts, each of which is rationalized by the whole. He asks us to imagine a person pushing a stone from point alpha to point omega. In order to reach point omega, he must first roll the stone to some intermediate point beta. If asked why he is pushing the stone to point beta, the man can truly respond "because I'm pushing it to omega," thereby showing that he is intentionally pushing it to beta. We can then repeat this procedure by picking a halfway point between alpha and beta, and do this infinitely many times. On Thompson's view, this shows that there is no foundational part of the process that serves only as means and not at the same time as an end to some more basic part.

To this argument, we might object that this way of classifying actions as intentional stretches our ordinary notion beyond recognition. It would be very unusual for an agent to have any of these infinitesimally small distances in mind as he strides from alpha to omega; if asked

about some arbitrary point, he would presumably be able to answer only after deducing that the point is practically relevant because it is between him and his destination. If he happened to miss hitting that particular point, perhaps because his path arcs a bit, he would be unlikely to view this is a mistake or a failure to take the means to his ends. Nor is an onlooker likely to be interested in the rationalizing explanation for these technical articulations, for there is no particular reason to be found for hitting each of these points. On the ordinary way of thinking about it, the agent is simply pushing the stone forward with the aim of reaching omega. If this is right, then we are not forced by this reasoning to give up on the notion of basic action.

The burden on those who accept that actions can be broken down into basic parts is to give an account of what basic actions are, and how they are related to more complex actions. We will canvas three possible positions here.

a. Bodily movements

The first candidate for basic action is the moving of one's body (we can call this 'Corporealism'). For this view to be plausible at all, we must understand 'bodily movement' very broadly so as to include effortful non-movement like "holding still," as well as whatever neural activity might be involved in purely mental actions like imagining a refreshing beverage. The central idea is that the only thing we can do directly or immediately is move our bodies, and all more complex actions are ultimately performed by way of moving ourselves.

The commitment to viewing the movement of one's body as foundational is implicit in the way many philosophers frame the topic at the outset. For example, the "problem of action" is frequently introduced by quoting Ludwig Wittgenstein, who asked: "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?" Wittgenstein, I should emphasize, did not actually think that this was a good question; he meant it to be an example of a mistaken way to think about the topic. However, many philosophers before and after have embraced this way of formulating the problem. And this formulation invites us to see the problem as a matter of understanding how bodily movements can be made into intentional actions with the addition of some extra feature. Similarly, Harry Frankfurt characterizes the contrast between what an agent does and what merely happens to him as the contrast "between the bodily movements that he makes and those that occur without his making them." Many more such examples can be found.

To motivate the view, the Corporealist can point to the difficulty of identifying cases in which we intentionally effect change in the world without using our bodies in any way (assuming, again, that bodily movement can include neural activity). Paradigmatic cases of action involve the exploitation of causal connections between the way we move our bodies and the results we ultimately desire. Even the most far-reaching political revolutions are accomplished by raising our voices in speech, moving our feet in protest, and standing in line at the voting precinct. To take a more mundane case, creating an omelet requires the agent to turn the dial of the stove with her hand to light it, whisk the eggs and milk together with a brisk motion of her wrist, place the pan on the stove so that it will heat, and pour the eggs into the pan so that they will cook. This is not to

say that she must fully grasp the nature of those causal connections, but in making these interventions, she presumes that they are there.

In contrast, we do not normally move our bodies by intentionally doing anything else, and there is nothing we do with the idea of causing our own bodies to move – we simply move them. One can, of course, grasp the fingers on one's left hand with one's right hand and manipulate them, but this is not the ordinary way in which we move our fingers. To be clear, the claim here should not be understood as denying that our actions frequently extend far beyond the limits of our bodies (more on this in section 4). Rather, it is that the primitive form of action is moving one's body, while everything else that we do is done by way of moving ourselves.

The student of neuroanatomy might point out here that our bodies move only as a result of events that occur in our brains, which then send signals to our muscles to contract in the right way. There is thus a sense in which the relevant causal chain begins before the body moves. However, this observation need not force us to conclude that our actions are at bottom neural events rather than bodily movements. The Corporealist is making a conceptual point, not a claim about what comes first in a chain of events. We do not cause our own bodies to move by way of intentionally initiating some prior neural event. We can see this by noting that the only way to intentionally initiate the right kind of neural event is to simply move one's body. If someone commands you to raise your arm, you can respond just by doing it. But if you are commanded to release acetylcholine – the neurotransmitter responsible for muscle activation – you cannot do it except by way of making some movement like raising your arm. Indeed, most of us have very little understanding of the anatomical underpinnings of the bodily movements we effortlessly make. It is therefore quite a strain to insist that such unfamiliar events are in fact the basic objects of agential control.

b. Volitions

Though predominant, the Corporealist view is not the only possibility on the table. A number of philosophers have held that basic actions are events that occur entirely within the mind, prior to any movement of the body. On this view, which we will call "Volitionalism," basic actions are mental events of *trying* or *willing* to do something. Any subsequent bodily movements and further outcomes that count as something the agent did do so in virtue of being the effects of her will. To be clear, although such mental events are also events in the brain, the Volitionalist is not primarily motivated by the "neural causes" argument just given. Our grip on the notions of trying or willing is meant to be independent of empirical neuroscience.

What does motivate the view is the idea that actions must be things that are fully under the agent's control. This intuition can lead us to conclude that the only thing we really control is our own will, or what we try to do. After all, it is at least conceptually possible to try to move one's body without anything happening – perhaps because, unbeknownst to you, someone has injected you with a powerful paralytic that prevents your limbs from moving. Still, *something* happened: you tried to raise your arm. In contrast, it is not clear that one can try to will an action and fail to do so. Thus, the only thing that is fully up to us is whether or not we will ourselves to act. The

Volitionalist concludes that acts of will are the basic actions, and that the bodily movements and all else that follows are the effects of our willings. As Jennifer Hornsby articulates this view, "Every action is an event of trying or attempting to act, and every attempt that is an action precedes and causes a contraction of muscles and a movement of the body (1980, 33)."

There is something to be said for the idea that our actions must be fully up to us and must therefore be limited to efforts that cannot fail. On the other hand, Volitionalism has the unhappy consequence of depicting us as relating to our bodies as a captain does to a ship, issuing orders to steer it that may or may not be followed. This portrayal belies the fact that we are essentially embodied creatures – not only thinking beings, but also animals. Distinguishing sharply between the agent and the rest of her body evokes an objectionable form of Dualism, albeit a dualism of brain and body rather than the traditional soul-body variety. Further, the idea that every action involves an event of trying to move one's body clashes with ordinary experience. Again, except in outré cases involving paralytic drugs, we do not experience ourselves as doing anything to cause our bodies to move. And as Gary Watson points out, the implication that all of our actions fundamentally occur within our bodies jars with the way that we normally think and talk about action: "The idea that your playing the piano, or hugging your friend, or cutting her hair, occurred inside your body seems not merely surprising but incredible (1982, 465)." Even basic actions seem at least sometimes to consist in making a change in the world and not just inside one's own mind.

c. Beyond the body

But once we entertain the thought that our actions almost always extend beyond our minds and into the world, why stop at the body? Why not simply say that in each case, the basic action is nothing less than whatever the agent intentionally accomplishes? Suppose Anton sets out to build a house and non-accidentally succeeds. Rather than saying he built the house by willing himself to do so, or by moving his body around in the requisite ways, could we not say that his action was nothing more fundamental than "building a house?"

To be clear, all the views in question will agree in this case that building a house was an action that Anton performed. The question is whether we can break this action down into a set of things that he did directly and the set of things that he accomplished by way of what he did directly. The Volitionalist and the Corporealist will both grant that Anton built the house in virtue of the fact that either his willings or his bodily movements caused the house to come into existence in just the way he envisioned it. Thus, what is at stake is not whether or not we are able to intentionally build houses, but whether or not building a house should be understood as a complex of more primitive actions and their effects.

Given this, it is not clear that our understanding of action is improved by claiming that everything we do intentionally is atomic, such that there is no distinction to be made between what we do directly and what we bring about. In fact, no one can avoid having to make this distinction eventually, even if we wish to draw the line in a place that is far removed from the contours of the body. Many of our actions ultimately aim at outcomes that are tenuous and far from guaranteed to happen. Perhaps Anton is building a house as part of an elaborate plan meant to culminate in

a coup against the Prime Minister of Malaysia. A lot will have to go right in order for the plan to succeed. But supposing that it does come off just as he intended, we surely do not wish to say that the most primitive action here was "deposing the Prime Minister," or that Anton's relationship to this outcome is just as direct as to his motions around the construction site. This shows that even a more capacious view of action does not avoid the need to draw the line somewhere between those things we do immediately and those we do by way of doing something else. We can grant that this line should not always be drawn at the limits of the body while remaining open to the idea that there is still some divisible structure within intentional action.

4. The Accordion Effect

Once we have committed to a view about what basic actions are, we face the further question of how basic actions are related to non-basic actions. One answer to this question is whimsically referred to as the 'Accordion Effect'. There is disagreement about how precisely to formulate the idea, but the central thought is that basic actions are related to non-basic actions by way of *causation*. As Davidson puts it, "an agent causes what his [basic] actions cause." That is, if event Y is the effect of some basic action X performed by the agent, then the causing of Y is also an action of hers – it is also something that she did. If Milgram moves her arm so as to push a button, and the effect of pushing the button is that an electric shock is delivered to a person in another room, shocking the person is something that Milgram did. In principle (though rarely in practice), Davidson's formulation allows the "accordion" of act-descriptions to be extended as far as the chain of cause and effect can be traced.

The phrase was coined by Joel Feinberg, who actually seems to have had a somewhat narrower claim in mind. He spoke specifically of the availability *in many cases* of a causative verb that can stand in for the complex phrase "A did X and thereby caused Y." For instance, rather than saying "A pushed the button and thereby caused person B to be shocked," we can felicitously say "A shocked B." But this will not always be true; Feinberg himself provides the example of "causing a person to laugh," which we cannot replace with a single transitive verb like "laughing him." For one thing, we tend to lack such a verb in cases where the causal upshot is another agent's action. Even though one agent can cause another to move out of the way by shouting "Move!" we would generally not say that the first agent moved the second, because this would seem to be incompatible with saying that the second agent moved herself. How precisely to reformulate the Accordion Effect to capture these exceptions is a matter of complexity. It will suffice here to note that if we like, we need not go in for the fully general Davidsonian version; we can allow for some exceptions to the spread of the accordion.

But even allowing for some exceptions, the Accordion Effect is sure to strike some readers as wildly over-inclusive. How could it be that our actions extend as far out into the world and the future as whatever they happen to cause? The crucial thing to understand here is that the Accordion Effect is not meant to capture only what we do *intentionally*, or what we mean to do. Rather, it is concerned with attributability: it aims to explicate when it is appropriate to attribute the occurrence of an outcome to a person and not just to some prior event. Many of the outcomes

that count as our actions under the Accordion Effect will end up being quite unintentional – effects we never anticipated or desired. Still, the thought is that those outcomes are attributable to us in virtue of being the effects of exerting our agency on the world. To determine which of the downstream events encompassed by the Accordion Effect are our intentional actions, we will need an answer to the question "What makes an event an intentional action?" This will be addressed in sections 6 and 7.

Finally, in emphasizing the importance of causation, we should not overlook the fact that the relation of constitution can also hold between basic and more complex actions. If Bas reposes in his armchair with the aim of having a relaxing evening, it is a strain to say that his posture is causing him to have a relaxing evening. Rather, his basic action constitutes a relaxing evening for him.

5. How many actions?

In some of the foregoing discussion, I have been begging a question which was a matter of lively debate in the 1970's and 80's, though it has since subsided somewhat. The debate concerns the numerical relationship between actions and their descriptions (both basic and non-basic). As I have set things up, when an agent acts, there is one action to which we can apply many different descriptions. The action will be intentional under some of these descriptions and not others. But there are those who disagree with this framework, holding instead that each distinct, true act-description picks out a distinct action. The disagreement is over how we should "individuate" actions. Imagine that Cheryl raises her arm, thereby doing the following things: 'voting', 'voting silently', 'voting in favor of the motion', 'casting the decisive vote in favor of the motion', and 'defeating John's bid for department chair'. Has Cheryl performed one action or six? Of course, she did six different *types* of things, but was there more than one *doing*?

According to the proponents of "coarse-grained individuation," she has performed only one action which can be correctly described in a variety of ways. In other words, her raising her arm is identical with her voting, voting silently, and so forth. This is the view that both Davidson and G.E.M. Anscombe held and the one I have been assuming. But according to the proponent of "fine-grained individuation," Cheryl has performed six distinct actions. The thought is that any two descriptions that make reference to different properties must correspond to different acts. Thus, even 'voting' and 'voting silently' are two different things that Cheryl did, albeit simultaneously, since silence is a property attributed only in the second description and not the first. The chief argument for embracing the fine-grained approach makes appeal to Leibniz's Law, which states that if X and Y are numerically identical, then X must have all and only the properties that Y has. Thus, if 'voting' and 'voting silently' do not share all the same properties, they must be different actions.

However, it is unclear that this is a context in which it is proper to apply Leibniz's Law. As Anscombe points out, it is not as though there was a voting that took place loudly in addition to the silent voting; the only vote that actually occurred was in fact silent. Thus, the only competitor to the voting which was in fact silent is a kind of abstraction, something that can "take place"

without any determination of volume or any other aspect of how it is done. The proponent of coarse-grained individuation can plausibly deny that Leibniz's Law applies to abstract objects of this sort.

Another issue which might seem to decide between the coarse- and fine-grained views concerns temporality. According to the coarse-grained theorist who subscribes to the Accordion Effect, some of the descriptions under which our actions are intentional will include the consequences of those actions, and those will be at a temporal remove from the original event. For instance, suppose that Raja moves his finger in such a way as to pull the trigger of a gun while aiming it at Eric, who later dies from the gunshot wound. In virtue of these events, it is the case that Raja killed Eric. However, the event of Raja's pulling the trigger occurred at an earlier time than the event of Eric's death. This suggests that the killing occurred later than the trigger-pulling, and must therefore be a distinct act, as the fine-grained theorist maintains. After all, if the killing is identical with the trigger-pulling, then we can truly say things like "the act of killing Eric caused the gun to go off," which sounds bizarre.

The coarse-grained theorist will deny, however, that the killing occurred later than the pulling of the trigger. The *death* certainly occurred later, but a killing is not the same as a death; a killing is an act that has a death as its consequence. It may be indeterminate at the time of pulling the trigger whether it is an act of killing; we must wait and see what happens. But if Eric does die, then the pulling of the trigger was also a killing. Thus, it is not in fact false, but merely odd, to say that the act of killing Eric caused the gun to go off. It is odd because we are using a description that had not yet come true at the time of the gun going off to refer to the cause of its going off. Anscombe likens this to saying "The widow stuck a knife into her husband," which is a similarly confusing way of saying "The woman stabbed her husband, thereby widowing herself." The coarse-grained theory is further bolstered by the observation that neither Raja nor the widow needed to do anything further once they used their weapons – their contribution to the chain of events was over at that point.

Debate over act-individuation has lessened because theorists came to agree, I think, that one's choice on this matter has few significant consequences for other questions of interest to action theory. I bring it up here in case this conclusion is mistaken.

6. The Causal Theory of Action

Let us take stock. So far, we have converged on the following ideas: (1) Actions are (broadly speaking) a kind of event. (2) They are intentional or not only under a particular description. (3) To be an action, an event must be intentional under at least one description. (4) Some actions are basic, while others are performed indirectly, in virtue of a basic action the agent performed. We are now in a position to ask the central question in the metaphysics of action, namely, "In virtue of what is an event an intentional action?" We will start by examining the historically dominant view on this question, which is known as the Causal Theory of Action. As the dominant view, the Causal Theory has been subjected to a variety of attacks, but it is not always made clear exactly what the alternative is. In the following section, we will canvass a few possibilities.

First, it is crucial to emphasize that the Causal Theory of Action is not exactly the same thing as the Causal Theory of Action Explanation that we discussed in Chapter Three. The latter is a view about how reasons explain intentional actions, while the former is a view about what makes it the case that some event is an intentional action. The two theories are certainly complementary, and many philosophers subscribe to both. It is possible to accept one but not the other, however.

The central claim of the Causal Theory of Action is that whether an event is an intentional action or not depends not only on how the event is described, but also on how it was caused. In other words, its status as an intentional action is not fully determined by its intrinsic properties; two events could share all the same intrinsic properties but differ in whether or not they are actions. As an analogy, think about what it is to have a sunburn. One must have an inflamed patch of skin, but this is not enough on its own, since the inflammation could have been caused by a variety of things. It is only a sunburn if the inflammation was in fact caused by the sun.

A more general philosophical analogy is the Causal Theory of Perception, which is an answer to the question 'In virtue of what is it the case that subject S perceives object O?'. According to this kind of view, a necessary condition of perceiving some object is that O play a causal role in producing S's visual experience. The thought is that it is not enough to count as genuinely perceiving a goat that it *seems* to you that you see a goat – even if in fact, there really is a goat in the vicinity. After all, your experience of seeing a goat could be a mere coincidence or contrivance; you could be looking at a sheep in dim lighting, or a hologram of a goat, without realizing it. To rule out such cases, the Causal Theory of Perception adds that your visual experience of a goat must have been caused, in part, by the actual goat. In general, if perception is to be a way of informing ourselves about an independently-existing world, then our perceptual capacities must be linked in the right way to that world. According to the Causal Theory, the "right way" should be understood in terms of causal dependence.

Similarly, agency is a matter of effecting change in the world, and that change must be linked in the right way to certain properties of the agent. It cannot be a mere coincidence that the agent's body happened to move around in a certain way, or that the world ends up just as she wants it to be. According to the Causal Theory of Action, we rule out this kind of coincidence by requiring intentional actions to have the right kind of causal history. Different versions of the theory will disagree about precisely what the right kind of causal history is, but all such approaches are united by the claim that having the right kind of cause is at least necessary for being an intentional action.

If we include the category of voluntary action as well as the more recently popular notion of intentional action, the Causal Theory has a very long history indeed, tracing back at least to Aristotle. Aristotle held that an action is involuntary if it is done out of ignorance or from compulsion, where an act is compulsory if its "moving principle" is external to the agent. For example, if you fall over only because another person pushed you, your falling is involuntary because the cause is outside of you and you "contribute nothing." A voluntary action, in contrast, requires that the principle "that moves the instrumental parts of the body" is internal to the agent. Of course, not just any internal cause will be of the right kind; a seizure is usually caused by some

internal condition rather than some external force, but is nonetheless involuntary. But the idea is that part of the task of action theory is to identify the further features a cause must have, in addition to being internal, in order to be of the right kind.

Versions of the Causal Theory were held by many throughout the Medieval and Modern periods of Western philosophy. Numerous versions took the form of holding that agents are distinguished by the possession of a *will* which issues in volitions. Volitions and their consequences, in turn, were thought to be voluntary actions in virtue of being the effects of the will. St. Thomas Aquinas, René Descartes, John Locke, David Hume, Thomas Reid, and John Stuart Mill each subscribed to a variety of this Volitional view, on which whether or not some occurrence was an action depends upon whether the will was its cause. Thomas Hobbes, William James, and Bertrand Russell rejected the notion of volition but also held causal theories, replacing volition in the causal history of action with imagination, appetite, and aversion. Even Occasionalists such as Nicolas Malebranche and George Berkeley thought that voluntary actions must have the right kind of cause, though they held that the cause must be God rather than anything internal to the agent herself.

As mentioned in Chapter Two, contemporary defenders of the view have shifted their attention away from accounting for voluntariness and are instead interested in distinguishing intentional actions from mere happenings. Further, the consensus about the relevant causes has shifted. It is now standard to appeal to the manifestations of the mental states of desire, belief, and (in some cases) intentions as the relevant causes. The chief idea – familiar from the Causal Theory of Action Explanation – is that intentional actions are the product of the agent wanting some outcome and believing that the action is a way to get it. For instance, if Annie intentionally sabotaged her roommate's job interview, she must have had some desire – that her roommate not move out, say – and a belief that spoiling the interview would further that outcome. Not only this, but her desires and beliefs must cause her to sabotage her roommate. Otherwise, the fact that she managed to ruin her roommate's prospects would be a mere coincidence rather than a consequence of her wanting that outcome. Some proponents of the Causal Theory argue that we should add the mental state of intention to the story, holding that the agent's beliefs and desires cause the acquisition of an intention to act, which then triggers and sustains the relevant actional mechanisms. Either way, the unifying commitment is that certain mental states of an agent must play a causal role in bringing an occurrence about if it is to count as an action. The contents of those states, in turn, will largely determine the descriptions under which the action counts as intentional. The relevant descriptions will be those that reflect what the agent desired and believed in so acting.

We can now put these insights together with the framework of basic and non-basic action. The Causal Theorist holds that our basic actions consist in some event – a volition, a bodily movement, or some more complex occurrence – that is caused "immediately" or "directly" by the relevant mental states or faculties. As noted in section 3, the notion of causal immediacy here may have to be understood quite loosely, especially by the Corporealist. The important claim is that when the agent's beliefs, desires, and intentions cause a volition, bodily movement, etc., this causal

sequence is the agent's intentionally doing something, and there is no proper part of this causal sequence that constitutes an intentional action in itself.

The Accordion Effect can then be brought in to expand the reach of one's basic actions to the causal consequences of those actions. Of course, we will count only some of those downstream consequences as something the agent intentionally did. The most important criterion is what the agent had in mind in so acting: the outcomes she desired or intended to achieve, and believed she might achieve, in doing as she did. If Milgram pushed the button because she desired to shock Janak and believed she could do it by pushing the button, then shocking Janak is a strong candidate for being an intentional action of hers. We sometimes even count those outcomes the agent knew he would bring about but did not specifically desire or intend, especially if we wish to blame the agent for it (more on this in Chapter Five, section 7). We might say that Milgram shocked Janak intentionally if she merely pushed the button intentionally while knowing that it would have the effect of shocking him, even if that outcome was not one of her goals – she just wanted to win the prize, say. On the other hand, if Milgram had no idea that pressing the button in front of her would cause Janak to get an electric shock, we will generally conclude that she did not shock him intentionally.

Further, the occurrence of the effect should not be merely accidental or extremely lucky. To intentionally bring about some effect, we must generally be exercising some kind of skill or understanding of a reliable connection between our basic action and that further effect, such that it is proper to see the effect as being sufficiently under our control. If Milgram desires to shock Janak and manages to do so by pushing a button, her action of shocking Janak may still not count as intentional if she was presented with fifty identical buttons and merely guessed which one would deliver the shock.

There may be other conditions as well – even some that are not best understood in terms of the causes of the action. For all that has been said so far, having the right cause is only a necessary condition and not a sufficient condition for an action's being intentional. Not all possible conditions we could add would be compatible with the naturalistic ambitions of Causalism, but some might be; we could add, for instance, that the agent must know what she is doing under some description. Some philosophers have gone further, however, and defended causal theories that are meant to provide both necessary and sufficient conditions for intentional action. It is worth emphasizing that contrary to what appears to be popular belief as of the writing of this book, Davidson was not one of them:

> Can we somehow give conditions that are not only necessary, but also sufficient, for an action to be intentional, using only such concepts as those of belief, desire and cause? I think not … for a desire and a belief to explain an action in the right way, they must cause it in the right way, perhaps through a chain or process of reasoning that meets standards of rationality. I do not see how the right sort of causal process can be distinguished without, among other things, giving an account of how a decision is reached in the light of conflicting evidence and conflicting desires. I doubt whether it is possible to provide such an account at all, but certainly it cannot be done without using notions like evidence, or good reasons for believing, and these notions outrun those with which we began (1974, 232-3).

However, other philosophers have been more optimistic than Davidson on this question. For example Michael Smith suggests that we can rule out the wrong kinds of causes and thereby offer sufficient conditions for (intentional) action:

> … we establish whether the agent acts by seeing whether this bodily movement is caused and rationalized in the right kind of way by some desire he has that things be a certain way, and some belief he has that a basic action of his—specifically, his moving his body in the way under discussion—has a suitable chance of making things the way he desires them to be: that is, he has to have a relevant a means-end belief. In this way we rule out the possibility that the bodily movement is caused by an alien force such as nerves, or the actions of remote super-scientists who are controlling bodies as if they were puppets. If these further conditions are met, then, according to the standard story, that bodily movement is an action. If not, then it is not (2012, 387).

Both Davidson and Smith are referring to the possibility of solving a problem for Causalism that first came up in Chapter Three: the problem of deviant causal chains. So it is to this objection that we must turn next.

a. Objection 1: Deviant Causal Chains, Redux

To review what was said in Chapter Three, the challenge is that no matter what a particular theory specifies as the causal antecedents of action, cases can be devised in which those causes produce an occurrence that is not an intentional action. We have already examined a classic example from Davidson, so here we will consider another classic example from Frankfurt:

> For example : a man at a party intends to spill what is in his glass because he wants to signal his confederates to begin a robbery and he believes, in virtue of their prearrangements, that spilling what is in his glass will accomplish that; but all this leads the man to be very anxious, his anxiety makes his hand tremble, and so his glass spills. No matter what kinds of causal antecedents are designated as necessary and sufficient for the occurrence of an action, it is easy to show that causal antecedents of that kind may have as their effect an event that is manifestly not an action but a mere bodily movement. The spilling in the example given has among its causes a desire and a belief, which rationalise the man's spilling what is in his glass, but the spilling as it occurs is not an action (1978, 157).

This example is similar to Davidson's nervous climber in that it works by introducing a loss of control between the agent's intention and the movement it causes. But we can construct other cases in which the deviance occurs at a remove from the agent's body. For instance, a third classic example is owed to Daniel Bennett, who asks us to imagine that a man intends to kill someone by shooting at him. His shot misses the intended victim by a mile, but it causes a herd of wild pigs to stampede and trample the victim to death. Here, the deviance lies between the means the agent had in mind and the way the death actually came about. Following Myles Brand, we can call

Bennett's example an instance of "consequential waywardness" to distinguish it from the "antecedential waywardness" exhibited in the first two cases.

The problem of causal deviance is a major challenge to the claim that having the right mental states as causal antecedents is sufficient for being an action. As noted above, this problem is one of the reasons why Davidson ultimately gave up on offering such an account, settling instead for a more modest view that takes having the right causal antecedents to be a merely necessary condition. Those who aim to defend the more ambitious view are committed to thinking that the problem of causal deviance can be solved.

There is good reason for optimism about solving the problem of consequential waywardness. One popular strategy appeals to the agent's "action plan." The thought is that when we set out to accomplish a goal that requires us to take some means, we must have a more-or-less specific plan as to *how* the means will bring about the end. These plans will rarely specify every detail, but they will embody a broad conception of the connection between the steps the agent is taking and the goal she aims to realize. For instance, Bennett's murderer presumably plans to kill his enemy by shooting him, such that he dies of a gunshot wound rather than a stampede of pigs. He likely does not have a plan that includes the exact trajectory of the bullet, or the exact time of death, but what he has in mind does effectively exclude a variety of deviant routes by which the enemy's death could be caused.

This is not to say that action-plans suffice to solve the problem without further work. Plausibly, there will need to be a condition ensuring that the match between the agent's action-plan and what happens is not accidental. The action-plan must in some sense guide what the agent does, or else the correspondence between plan and result could itself be the product of consequential deviance. Moreover, we might worry about whether our action-plans will always be detailed enough to exclude deviant cases. Still, the broad consensus is that this kind of case is far less troubling than cases of antecedential waywardness.

In order to rule out causal deviance antecedent to the basic action, we must specify what it is for a mental state or event to cause an action "in the right way." Some philosophers have suggested that this problem is ultimately empirical rather than philosophical. What is needed, we might think, is a completed neurophysiological story of the normal causal route between intention and action that can differentiate between controlled movement guided by an intention and uncontrolled movement caused by some intervening factor. But while this is a tempting thought, it also begs the question in this context to simply assume that there will be such a story to tell. It would be better for the Causal Theorist to offer at least some philosophical guidelines for such an account.

One strategy is to stipulate that the relevant mental items – belief/desire pairs, intentions, or their manifestations – must be the proximate cause of the action, excluding any intermediate events like nervous spasms. The thought is that this leaves no room for waywardness in the causal chain leading up to the action. Read strictly, this strategy is off limits to the Corporealist, since it is simply false that the neural realizers of our beliefs, desires, and intentions are the proximate causes of our bodily movements. Rather, signals must be sent from the brain to the limbs, acetylcholine released, muscles activated, and so forth. To employ the proximate-cause strategy,

then, the Corporealist must include this entire neurophysiological chain as part of the action. Perhaps this is acceptable in itself, if counterintuitive. However, it has the implication that causal deviance could enter the chain after it has been initiated but before the relevant bodily movement has been completed. And deviance at that point cannot plausibly be dealt with using the "action-plan" strategy for consequential waywardness, since most agents have no conception at all of how their intentions cause their bodies to move. Thus, this strategy is perhaps best suited for the Volitionalist to employ.

A second strategy requires that intentional actions be "sensitive" to the content of the agent's intention, where sensitivity is understood counterfactually. In the cases of Davidson's climber and Frankfurt's accomplice, it is a mere coincidence that the spasmic result of their nerves happened to match what they intended to do. Any other intention that made them equally nervous would have had the same result. In non-deviant cases, by contrast, there is counterfactual or explanatory dependence between intention and action: if the agent's intention had been different, her behavior would have been different as well. We expect this kind of counterfactual dependence because in non-deviant cases, it is the intention that at least partially explains the resulting action, rather than some intervening factor that screens off the intention.

A third strategy (potentially compatible with the second) is to claim that the intention must not only causally initiate the action, but also sustain and guide it. This strategy takes its cue from an argument offered by Frankfurt that was meant as a criticism of the Causal Theory. Frankfurt dismissed the Causal Theory as being concerned only with what happens before the action begins, suggesting instead that what differentiates actions from events is that the former are essentially under a person's guidance. If the person comes to believe that her action is not on course to achieve her goal, she is disposed to intervene and put it back on track. It is not clear why the Causal Theorist cannot simply take this insight on board, however, by offering an explanation of what it means for the person to guide her behavior in causal terms. A promising model here is the "negative feedback loop," in which sensory feedback is used to check whether a particular bodily movement matches the content of what was intended, and if not, a corrective motor command is issued. On this model, the causal contribution of the agent's mental states is not completed until the action is completed.

Fourth, one might seek to supplement the causal story by adding that the action must be caused in part by the fact that one has a rationalization for so acting. That is, perhaps it is not enough that the right combination of attitudes cause some behavior; the logical relations between the contents of those attitudes must also feature as a cause. Where most causal theories specify only that the agent's mental states must both cause and rationalize her actions, this approach holds that they must cause *by virtue of* rationalizing. Thus, Frankfurt's protagonist did not spill his drink intentionally because although there was the right kind of logical relation between his beliefs, desires, and the resulting spill, those attitudes did not cause the spilling by virtue of that relation. Rather, they caused it by virtue of their unnerving properties. The success of this strategy will depend on whether there is some special trouble in making sense of causation by rationalization, over and above our general trouble with understanding causation as such. Further, it is not clear that it is entirely compatible with the event-causal framework that Causal Theorists are usually

wedded to, since it assigns a causal role to the fact that certain logical relations hold, and facts are not events.

This is only a sampling of a vast literature which continues to grow. At the writing of this book, a consensus has not been reached among defenders of the causal theory as to the best approach, and many philosophers of action remain dubious that the problem can be solved. But it is worth noting here that the problem of deviant causal chains threatens not only causal theories of action, but causal theories of anything – perception, knowledge, mental content, and so forth. Insofar as this kind of theory has struck many thinkers as promising in a variety of domains, we might see this as reason to persevere in attempting to solve the challenge. That said, one could also see the history of failed attempts as inductive evidence that no solution is forthcoming.

b. Objection 2: The Disappearing Agent

The Causal Theory aims to locate actions within the overarching framework of cause and effect, as things that happen just as other kinds of events do. Some critics argue that the approach is hopeless, since a chain of cause and effect could never constitute an agent's doing something. This criticism is often referred to as the "disappearing agent" objection. Thomas Nagel expresses the worry most eloquently in *The View From Nowhere*:

> Something peculiar happens when we view action from an objective or external standpoint. Some of its most important features seem to vanish under the objective gaze. Actions seem no longer assignable to individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part … There seems no room for agency in a world of neural impulses, chemical reactions, and bone and muscle movements. Even if we add sensations, perceptions, and feelings we don't get action, or doing – there is only what happens (110-111).

Hornsby echoes this thought, arguing that "… agency cannot be portrayed in a picture containing only psychological states and occurrences and no agent making any difference to anything" (2004, 12) Critics conclude that no account on which actions are events within the causal order could succeed.

Again, this kind of objection lands most cleanly against the strongest version of the Causal Theory, on which having the right kind of causal history is not only a necessary but also a sufficient condition of acting intentionally. That view purports to be an analysis of what it is to act in terms only of events, and it is legitimate to worry that the resources it offers are too thin to capture the meaning of the concepts 'agent' and 'action'. Even then, however, the critic must be careful not to rest the objection on what is arguably just a bad metaphor. The complaint is sometimes put in terms of the subject being a "mere arena" in which psychological states are contained, such that she is not involved in the interactions between mind and body. But the Causal Theorist is in no way committed to this way of thinking about the relationship between the subject and her own mind. Indeed, this seems to be a prime example of a category mistake: "I see that there are mental

states, and a body that moves around in virtue of this mental activity, but where is the person that does the moving?"

Further, there are versions of the Causal Theory that take the objection seriously but claim that it can be solved by a view with the right structure. J. David Velleman is one of those who originally deployed this objection against what he viewed as overly simplistic versions of the approach, but he argues for a solution that is itself a version of the Causal Theory. On his view, one of the psychological antecedents of action must be no ordinary desire, belief, or intention; it must be a mental state that, as he puts it, "plays the functional role of the agent." That is, it must be a state that mediates between the agent's reasons and the forming of an intention in light of those reasons, and that guides the agent's bodily movements in the executing of that intention. The identification of the agent with a particular attitude that plays the relevant causal role helps to ensure that even if we view ourselves from the external standpoint, in terms of cause and effect, the action will be attributable to a source that we can legitimately view as an agent.

Alternatively, as we will see in more detail in Chapter Eight, Frankfurt articulates a view that is potentially compatible with the Causal Theory and that identifies the agent with a complex psychological structure composed of the person's wholehearted higher-order desires and loving commitments. On this view, there will not be a single attitude that "plays the role of the agent" whenever the person acts, but there will be some guiding attitude that is an element of the complex that constitutes the agent and thus plays the relevant role in that instance.

Finally, for those Causal Theorists who are interested only in an ontological and not a conceptual reduction (see Chapter Two, section 12), the objection may seem less worrisome. *Of course* the account does not contain agents – the whole ambition is to avoid adding agents to the basic furniture of the world. Philosophers with this more modest aspiration may well grant the critic that for most of our purposes, we cannot simply replace the use of the concepts 'agent' and 'action' with talk of causation.

7. Alternatives to the Causal Theory

As the "orthodox" view, the Causal Theory of Action is often the target of criticism. I think it is fair to say, though, that its opponents are not always clear about what exactly they propose to put in its place. The following are some alternative conceptions we might have as to what makes some occurrence an intentional action.

a. Quietism

One alternative is simply to deny that there can be any account of what makes something an intentional action. This claim might be motivated by a further thesis, which is that 'intentional action' should not be understood as a species of some more general category of 'action'. If we start by accepting this genus-species taxonomy, then it can seem inevitable that there must be some "extra feature" in virtue of which some generic actions count as intentional actions. This is the conclusion we are led to if we embrace the "subtraction" model of the problem of action: "What

is left over if we subtract the fact that my arm went up from the fact that I raised my arm?" But it is possible to reject this formulation, holding instead that 'intentional action' is not divisible into 'actions' plus some 'extra feature'. This means that the most a philosophical account of intentional action can do is point to characteristics that intentional actions tend to have. This would be to give some elucidation of the concept 'intentional', but it would not tell us anything further about how some things in the world come to satisfy that concept.

This may well be the place we should end up, and we should be wary of helping ourselves to the starting assumption that there must be an "extra feature" that will serve as an answer to this question. On the other hand, to conclude that intentional action is a *sui generis* phenomenon of which no further metaphysical account can be given would be something of a philosophical disappointment. Thus, other attempts to give such an account should be pursued as far as we can take them.

b. Agent-causation and causal powers

The Causal Theory of Action takes a naturalistic approach to understanding what actions are, in that the conception of causation it appeals to is the same kind that is at issue in science. What this conception is remains something of an open question. Davidson was operating with the "nomological" model advocated by Carl Hempel, according to which causation is a relation between two events. On this model, for C to be the cause of E, there must be some law-like connection that holds between events of these types, such that given certain background conditions, the occurrence of an event of type C is sufficient for the occurrence of an event of type E. However, the nomological model is not the only conception of causation available. We can also think of it in terms of a structure that holds between variables, representing counterfactual dependence. Roughly, if E causally depends on C, then E would not have occurred (or would have been less likely to occur) if C had not occurred, and E would occur (or would be more likely to occur) if C did occur. "Interventionist" accounts of causation emphasize the way in which these counterfactual relationships can be revealed by manipulating the variables and isolating the changes that result.

Either way, an alternative to what we have been calling the Causal Theory rejects this broadly scientific understanding of causation. Instead, it appeals to a distinct conception of "agent-causation," on which causation is a primitive relation between an event and an *agent* (see also Chapter Three, section 4b). This conception denies that the relevant cause at work in claims like "Steve shook his fist" is some event – the occurrence of an urge to shake his fist, say – or any other aspect or property of Steve. It also denies that any antecedent event fully caused Steve to decide to shake his fist, such as the behavior of the kids these days. Rather, nothing less than Steve himself is the cause of fist's shaking, and his decision to shake his fist was not the effect of any prior cause. As an agent, on this view, he is able to originate causal chains. Of course, many event-causal theorists will wish to agree that there is a sense in which agents cause things, but what they mean is that where there is action, there is a constellation of mental and physical events that amounts to

an agent's causing an outcome. The agent-causal theorist rejects this reductive analysis, holding instead that agent-causation is primitive.

This general type of view is sometimes developed more specifically to refer to a collection of "causal powers" that are possessed by an agent. Powers are a kind of disposition, or tendency to manifest certain properties under certain conditions. For example, fragility is the disposition to break when struck. Most powers are understood to be "one-way," in the sense that they are tied to a specific manifestation and not to the absence of that manifestation. Fragility is not the disposition to either break when struck or not; rather, the tendency not to break when struck – robustness – is a distinct disposition. But some philosophers have claimed that agency involves the possessions of a set of "two-way powers," which means that if an agent has the power to perform act A at time *t*, then he also has the power not to perform A at *t*. This is a way of capturing the idea that it is up to Steve whether to shake his fist; it is not simply a brute disposition that he is guaranteed to manifest when certain circumstances obtain. Because Steve is an agent, he can make things happen by exercising his two-way causal powers.

It is not clear that the agent-causal-powers approach is compatible with a naturalistic, scientific understanding of the world. The view has traditionally been associated with mind-body Dualism, since agent-causal powers were historically attributed only to immaterial souls. That said, more recent defenses avoid the commitment to Dualism, holding instead that such powers are shared by non-human animals or even inanimate objects. Still, the idea that agents are able to originate causal chains, and that the power to act is always accompanied by the power not to act, seems to be incompatible with the deterministic picture with which science presents us. If determinism is true, then every physical event – including the bodily movements of agents – is necessitated by some prior event, together with the laws of nature and the background conditions. And although it is possible to deny the truth of determinism, our best scientific evidence suggests that it is true at the non-quantum level.

Further, the agent-causal approach has been charged with failing to be explanatory. Whereas the event-causal framework promises to shed light on why an action occurred by identifying the events that led to the agent doing as she did, the agent-causalist seems to have little more to offer as an explanation than "The act occurred because the agent exercised her power to act." This kind of claim is reminiscent of Molière's parody of a non-explanation in *Le Malade Imaginaire*, in which the fact that opium induces sleep is explained by appeal to its "dormitive powers." The problem with this "explanation" is that we have no grip on what a "dormitive power" is except by reference to its putative effect: inducing sleep. Similarly, the defender of agent-causal powers must explain what kind of independent grip we have on such powers, other than by reference to the production of the very actions we are hoping to explain.

c. Formal causation

A third alternative to the Causal Theory rejects altogether the idea that the antecedents of an event determine whether or not it is an intentional action. Instead, it holds that certain events or processes are intentional actions in virtue of their *form*. It is not the case that the very same

occurrence could either be an intentional action or not, depending on some other factor, since intentional actions display a kind of unity that an unintentional event could not have. After all, nearly every intentional action we are interested in takes time to be completed; it is a vast oversimplification to think of them as momentary events. Rather, they are often a series of events: adding flour to the bowl, then water, then yeast, then salt. It is only when such events fit together in the right way that they can properly be described as the intentional action of making bread.

In virtue of what does an occurrence or series of events have the form of an intentional action? One idea is that the agent herself must bring the occurrence under the relevant concept by thinking of what she is doing in that way. Shantha is intentionally making bread if that is the description she has in mind, such that she sees each step as part of this teleologically-structured whole. If she does not think of these steps as part of making bread, then they are merely a disunified collection. Perhaps her hand is simply spasming and accidentally knocking one thing into the bowl after another with no further reason or goal. But if she does conceive of what she is doing as making bread, then this is sufficient to transform an otherwise disparate series into a unified whole with the form of an intentional bread-making. We can call this a matter of "formal causation," following Aristotle's taxonomy of the four causes (material, efficient, formal, and final), though it is important to remember that this kind of causation has nothing to do with mechanics of the physical motions involved in acting.

A Kantian version of this approach emphasizes not only the agent's conception of what she is doing, but why. As Christine Korsgaard characterizes this kind of view, the form of an action lies in the way that it is chosen by the agent. The agent chooses not only to perform some act, but to perform it for the sake of some end: to make bread in order to please one's mother-in-law, say. Further, to be an action, this package of act and end must be chosen in light of the correct principles. According to the Kantian, these are the Categorical and Hypothetical Imperatives (more on this in Chapter Seven, section 3). Taken together, the action embodies a maxim: "This act, done for the sake of this end, at the right time, in the right way, is choiceworthy."

These approaches require that whatever structure is essential to the form of intentional action be found in the agent's conception of what she is doing. Not only must she *have* a concrete conception of what she is up to, but she must also be in a position to distinguish between the parts of her action that are undertaken as means to end and the parts that are mere side effects (insofar as we think there is a fact of the matter about this). For instance, Shantha might in part conceive of what she is doing as "using up all the yeast," though this is a side effect and not part of her goal. Thus, her action of using up all the yeast must be represented differently from the actions that are the means to her end of making bread. This is not necessarily an objection in itself, but it is a challenge for this approach to make good on the idea that whatever structure we actually find in intentional action will be present in the agent's conception of what she is doing. Further, we may not wish to grant the agent unlimited authority to determine the fact of the matter. Watching Shantha pour out the whole package of yeast even though it is more than she strictly needed, we might conclude that she really does aim to use it up (so that her mother-in-law can't have any), even if she herself does not conceive of her action in that way.

d. An "actish" phenomenal quality

As noted in Chapter Two, there are those who have maintained that intentional action necessarily involves a certain kind of phenomenology. Following Carl Ginet, we can call this an "actish phenomenal quality." Ginet describes the feeling as "I directly make it happen;" others have characterized it in terms of being in control, or being the author of the action. In addition to taking this kind of agentive experience to be necessary for action, it is possible to hold that it is definitional of it. Ginet himself is a Volitionalist and argues that "The actish quality of the mental occurrence is enough in itself to make the occurrence a mental act" (1990, 14). This approach to understanding what intentional action is prioritizes the deliverances of introspection, holding that we can really only grasp the concept of an action from the inside by consulting our own experience.

Though it has philosophical defenders, the experiential conception of agency often crops up in the context of research in psychology and neuroscience that aims to debunk the idea that we are agents. The strategy is to identify conditions in which the conscious experience of agency is absent, but that otherwise have the features (like goal-directedness) we usually attribute to exercises of agency. The skeptics then conclude that agency is an illusion, since the concept of agency has a necessary condition – the conscious experience of acting – that our goal-directed behavior does not meet. For example, Daniel Wegner points to conditions like Anarchic Hand Syndrome, in which the hand engages in behavior like buttoning up a shirt or grabbing food off a plate, but where the owner of the hand does not feel as though he is controlling it. Wegner concludes that the feeling of conscious willing is disconnected in general from the production of "action," such that it turns out that our "actions" in fact happen to us. But this form of argument is only plausible (if at all) if we insist that all actions, even pathological ones, must manifest the actish phenomenal quality. There is thus some reason to avoid committing to this strong claim.

8. Omissions

In addition to the various interventions we make in the world, we sometimes omit to act. My plants might die because I omitted to water them, or my friend might be angry with me because I omitted to pick her up at the airport. Further, although many omissions are unwitting or accidental, some of them are intentional: perhaps I intentionally omit to water my friend's plants while she is away as revenge for some previous slight. How do intentional omissions fit into the frameworks of agency we have considered so far?

There are some cases in which an omission is simply an intentional action under a privative description. For instance, one might do something in order to prevent oneself from acting, or in order to restrain oneself. An agent who is sorely tempted to reach out and touch the artworks in a sculpture garden might intentionally put her hands in her pockets to ensure that she does not succumb to the urge. This kind of case poses no special problem for the views discussed in this chapter, since her omitting to touch the sculptures just is her sticking her hands in her pockets under a different description.

However, many cases do not involve this kind of active refraining. It is not as though my omitting to water my friend's plants consists in my standing in front of them and restraining myself from making any pouring motions. Rather, most likely, I am simply doing something else entirely at the time I should have watered the plants. We might try to extend the strategy just mentioned to include these kinds of cases by claiming that whatever it is I was doing at that time – having coffee with a student, say – was also an action of not watering the plants. Unlike in the first case, I am not having coffee in order not to water the plants. Still, under some conditions, perhaps the event of my having coffee is also intentional under the description 'not watering my plants'. If I deliberated about whether to water the plants and decided to go for coffee instead, this might make such a description applicable. "Here I am, not watering Emily's plants!" I might whisper to myself at the coffee shop with a spiteful chuckle.

Alternatively, we might characterize omissions as absences of action rather than as actions under an omissive description. When I omit to water Emily's plants, I simply did not act. This approach avoids the challenge of trying to find an appropriate action in the vicinity whenever there is an omission, and fits more naturally with how we normally talk about omissions. It also renders most omissions outside the purview of the philosophy of action, though we might still be interested in the many similarities between actions and intentional omissions. This view encounters a challenge, however, in explaining how omissions can stand in causal relations to anything. It seems as though my omitting to water the plants caused them to die, but at least some have found it puzzling to think that an absence of action can cause anything. Further, if the omission was intentional but uncaused (supposing that absences cannot be caused), then it is presumably not intentional in virtue of its causal antecedents. This does not contradict the Causal Theory of Action, if omissions are not actions, but it does cast some doubt on the explanatory power of that theory if it fails to give a unified explanation of all things intentional.

9. Mental actions

The diet of examples we are given in the standard literature in action theory is heavy on cooking, manual labor, crushed toes, and violent deaths. It rarely includes what we might want to describe as "purely mental" actions: imagining a winged horse, silently counting to ten, attending to the blaring siren, or making a difficult decision about what to do. Are there such things as genuinely mental actions? Of course, if Volitionalism is true, then all actions are fundamentally mental acts of the will. Even so, we can still ask whether there are any mental actions other than volitions.

Common sense suggests that we perform mental actions all the time, but there are theoretical grounds for skepticism. For one thing, if Corporealism is the right account of basic action, this may be hard to reconcile with the idea that some actions involve no bodily movement. Davidson gestures at a definition of bodily movement that is meant to allow for mental acts, but does not elaborate on what this account might look like. As long as Physicalism is true and mental acts are neural events, however, we might understand bodily movement in such a way as to include certain patterns of neural activity.

More worrisomely, when it comes to mental events like thinking a certain thought, judging that P, or deciding to A, a problematic circularity appears to arise. Ordinarily, in order to perform some action A, one must first have it in mind to do A – one must intend to A, or otherwise be trying to A, or at least believe that A-ing might be the result of something else one is trying to do. But when it comes to thinking, judging, or deciding, one cannot have the relevant action in mind without actually performing the action. That is, if intentionally thinking the thought "The Burj Khalifa is in Dubai" requires that one first bring that content to mind in order to think it, the result will have been achieved before any further act can be performed. Similarly, to intentionally judge that P, it seems that one will first have to identify P as the proposition to be judged, and to decide to A, one must first identify A as the action to be decided on. "I know – I'll judge that the Burj Khalifa is in Dubai, and I'll decide to go to Dubai in order to see it," you might think to yourself. But this would be already to have made the relevant judgment and decision. Galen Strawson concludes that most of our mental agency is limited to creating conditions that are hospitable to the relevant contents coming to mind, and is in that sense "prefatory" or "catalytic" rather than direct.

One strategy for avoiding this kind of skepticism is to reject the conception of agency that requires the agent to have advance awareness of the act she is trying to perform. Thinking, imagining, judging, and deciding are all things that can be done in light of reasons, just as bodily actions are. We might think it enough that we have a kind of "evaluative control" over these mental events, in the sense that they are responsive to our assessment of what we have reason to think and do. Alternatively, the agency involved in judging and deciding might be viewed not as a matter of producing those events, but rather as a kind of ongoing commitment we undertake to stand behind our assessments of what is true or worth doing and take responsibility for them. In any case, this is a topic that deserves more attention than it often gets.

Summary

We have covered a large amount of territory in this chapter; indeed, it would have been possible to write a whole book just on these issues. We started with the idea that actions are a kind of event, though they are almost always dynamic events that take time to be completed. Second, whether or not any particular event is an intentional action depends on how it is described – an event is an intentional action only "under a description." On Davidson's way of thinking about it, if there is at least one description under which an event was an intentional action, then it is an action under every applicable description. This claim allows us to say that George's putting his sweater on backwards was an action of his, since he did intentionally put it on even though he didn't intentionally put it on backwards. Further, though there is some disagreement about this, the predominant view is that "putting his sweater on" and "putting his sweater on backwards" were not two different actions George performed; rather, there is only one action (the event) to which we can apply a variety of descriptions.

Next, we canvassed three different views about what "basic" actions are as well as the skeptical viewpoint that there is no such thing as basic action. Traditionally, the main contenders

for basic action have been volitions – mental events of willing or trying – and bodily movements. Proponents of these views will hold that all the agent ever does directly is will something to happen, or move her body, and that everything else she does is accomplished by way of being an effect of these foundational acts. A third view voices skepticism about the idea that the sphere of direct agency is limited to our minds or bodies, holding instead that nothing less than the entire intentional action should be considered atomic.

But in virtue of what is an event correctly described as an intentional action? According to the Causal Theory of Action, it is at least a necessary condition that the event have the right kind of causal history. It must have been the product of the agent's will or caused by a combination of her beliefs, desires, and intentions. On some versions of the view, having the right kind of causal history is also sufficient for being an intentional action, although these views tend to pack other conditions into the causes – that the relevant beliefs and desires also rationalize the action, for instance. Either way, the basic insight is that we discover that some behavior wasn't really an intentional action by finding out that it was caused by a spasm, or a hypnotist, or aliens using brain-control devices. Actions must proceed from the right kind of source.

The Causal Theory is subject to recalcitrant objections, however, that might be avoided by embracing a different conception of action. Each of the four alternatives we considered has the virtue of being invulnerable to the problem of deviant causal chains. Three of them hold the causal history of an event to be irrelevant to whether it is an intentional action. That fact is either basic and unanalyzable, or lies in whether the event has the form of an intentional action, or in whether it has an actish phenomenal quality to it. And the fourth approach holds that actions must be the product of a special kind of agent-causation which was clearly not exercised in the deviant examples.

Finally, we noted that not everything we do intentionally consists in some bit of observable behavior. We sometimes intentionally omit to act, and some of our intentional actions are purely mental. Omissions and mental actions have distinctive features that pose trouble for assimilating them into some of the standard frameworks. If we think such cases should be dealt with by a theory of intentional action – though we might deny that they should – then we must be careful not to focus exclusively on overt behaviors in our investigations.

Suggested Reading

Donald Davidson's argument that actions are events can be found in "The Logic of Action Sentences." For vigorous dissent to this claim, see Maria Alvarez and John Hyman, "Agents and their Actions." The introduction to Jennifer Hornsby's book *Simple-Mindedness* usefully articulates the distinction between particular actions and "things done."

The idea that actions are intentional only under a description is defended both by Donald Davidson in "Actions, Reasons, Causes" and by G.E.M. Anscombe in Intention and in her paper "Under a Description." The notion of basic action is first developed in Arthur Danto's paper "Basic Actions." Skepticism about that notion is voiced by Michael Thompson in Part Two of *Life*

and Action, and well as by Douglas Lavin in "Must There Be Basic Action?" Hornsby's book *Actions* is a great example of the Volitionalist view, while Corporealism is argued for in Davidson's "Agency." Anton Ford's paper "The Province of Human Agency" defends the more capacious "Materialist" view and is a wonderful overview of the dialectic between these three approaches. The idea of the Accordion Effect can be found in Joel Feinberg's "Action and Responsibility" and is further discussed in Michael Bratman's "What is the Accordion Effect?"

The fine-grained approach to act-individuation is exemplified by Alvin Goldman's *A Theory of Human Action*, while the coarse-grained approach is found in both Davidson's and Anscombe's works. For more on the Causal Theory of Action, Michael Smith's "The Humean Theory of Motivation," "The Possibility of Philosophy of Action," and "Four Objections" provide an especially clear statement. See also Alfred Mele, *Springs of Action*; Goldman's *A Theory of Human Action*; Berent Enç, *How We Act*; J. David Velleman, *The Possibility of Practical Reason*; and Kieran Setiya, *Reasons Without Rationalism* and "Reasons and Causes."

Chapter Five of Erasmus Mayr's book *Understanding Human Agency* is enormously helpful in summarizing the debate over deviant causal chains. Some of the source literature includes John Bishop, *Natural Agency*; Myles Brand, "Intending and Acting," Mele's *Springs of Action*; and Christopher Peacocke, "Deviant Causal Chains."

Roderick Chisholm's book *Person and Object* offers an example of an agent-causal theory of action as traditionally understood, while Helen Steward's *A Metaphysics for Freedom* and Maria Alvarez's "Agency and Two-Way Powers" are good resources for exploring the "causal powers" proposal. The idea that intentional action is not an action or a movement with some "extra feature" is voiced in Ford's paper "Action and Generality." For a way of understanding how intentional action could be thought of as instantiating formal causation, see sections 1-2 of Sarah Paul, "Deviant Formal Causation."

Finally, Galen Strawson's paper "Mental Ballistics" is a landmark for articulating skepticism about mental action, and the collection *Mental Actions*, edited by Lucy O'Brien and Matthew Soteriou, contains a number of important discussions. Randolph Clarke's book *Omissions* is a good place to start on that topic.

5. Intention

So far, the focus has been on what it is for an action to *be intentional*, or to be *done intentionally*. But in addition to the adjectival and adverbial forms, 'intention' also functions as a noun. We commonly talk about doing something *with an intention*, or about our *intentions for the future*. We say things like "You brought up last Christmas with the intention of hurting my feelings!" or "I'm staying home for Christmas this year, but next year my intention is to spend it in Rotterdam with or without you." These uses suggest that intentions are things that we can have or states that we can be in. In this chapter, we will start by examining arguments for and against thinking about intentions as real mental items. Then, assuming that they are mental items, we will investigate what kind of mental item they might be.

1. Methodological priority: present or future?

Where we end up on the question of what it is to act with an intention, or to have an intention for the future, can depend significantly on where we start. As we saw in Chapter Two, when we set out to theorize about action, we must make a choice about which kinds of cases we take to be paradigmatic. In the context of thinking about intention, a critical choice concerns whether we begin by focusing on the present or on the future. Do we take as our paradigmatic examples actions that are in progress, or to be done now? Or do we begin with the observation that we frequently plan future actions in advance, especially those that are of the most significance to us? Following Michael Bratman, we can refer to the first approach as placing methodological priority on intention-in-action, while the second approach places methodological priority on future-directed intention.

Theorists who prioritize the present have tended to conclude that we can do without a substantive theory of intention. They suggest instead that we can reduce talk of intention to some other feature of intentional action. In his early work, Donald Davidson fell into this camp. He proposed that the term 'intention' does not refer to some mental state or property but is instead "syncategorematic:" it is a verbal mechanism we can use to redescribe an intentional action in terms of the reason for which the agent acted. "James went to church with the intention of pleasing his mother" is simply another way of saying "James's reason for going to church was to please his mother." This suggestion is pleasingly parsimonious, if nothing else, since it promises to reduce the notion of intention to the phenomenon of acting for a reason.

Davidson does not offer any real argument for this claim, and as we will see, later came to reject it. More recently, Michael Thompson has also proposed that talk of intention should be understood in terms of intentional action. He argues that intentions cannot be mental states, let alone propositional attitudes like belief and (some forms of) desire, because they are not static and do not take propositions as their objects (see also Chapter Three, section 4d). He points out that actions are processes that take time to complete, and that insofar as they are still intended, they are

essentially incomplete. We normally speak of intending *to do* something, not of intending that something be the case (although we should note that the latter is not unheard of, especially in situations where one agent has power over another, e.g. "I intend that you clean your room today, son!"). And in intending to do something, one's relation to the goal of the action is constantly evolving as the act-process unfolds, and hence not one of statically representing a proposition.

Thompson concludes that intending is not a psychological state at all, but rather a form of doing. To say that Dave intends to fly to San Francisco tomorrow is simply another way of saying that he *is* flying to San Francisco tomorrow: the act-process of traveling to the Bay is already unfolding. We speak of intending, Thompson suggests, especially when we wish to indicate that very little progress in the action has been made, such that it is awkward to say of Dave that he is currently flying to San Francisco when he is having dinner in New York. But he warns that we should not make the mistake of supposing that purportedly psychological explanations that appeal to intending or wanting are the fundamental form of explanation, since such explanations can always – with some stretching of common usage, perhaps – be translated into "naïve" explanations that appeal only to something else that the agent is doing.

Davidson later became convinced that this kind of deflationary analysis of intention was inadequate. Even if it is compatible with many cases, there seem to be cases left over in which there is an intention without any relevant intentional action, or even deliberation, to reduce it to. Call such cases "pure intending." The thought is that an agent might come to have an intention for the future – in Davidson's characteristically whimsical example, the intention to build a squirrel house one day – without ever having consciously deliberated about whether to build one or decided to do so. And this intention might in fact never lead to any kind of action; the agent might never even try to put nail to board. If we can nevertheless make sense of him having the intention in this kind of case, then it looks as though we cannot explain what is going on in terms of any intentional mental activity or action. Davidson concludes that intention must be some kind of attitude after all, and most subsequent theorists have agreed with this.

The case for acknowledging the existence of intentions need not rest on the mere possibility of an isolated pure intention. Bratman points out that we human beings are the kind of agents who frequently plan for the future. If we begin our theorizing with the idea that intentions are the building-blocks of plans and attempt to articulate the complex dynamics of planning, we will encounter little temptation to think that intentions can be explained away in terms of action. Instead, we will be in a better position to see the crucial role intention plays for us in coordinating our actions over time, both individually and with other people.

Thus, though not all philosophers of action agree that intentions are states that are distinct from intentional actions, the remainder of the chapter will assume that they are. But what kind of states are they?

2. Goal states and plan states

As a first step, we may note the close connection between an agent's intentions and her goals or ends. There is a broad sense of 'intend' in which it simply denotes the state of being settled

on a goal to pursue. This sense can be brought out most clearly by using the phrase "the intention with which one acts," as when we say "Sonia Sotomayor applied to Yale Law with the intention of becoming a federal judge one day." Some philosophers have argued that the word 'intention' is ambiguous, however, and that there is a narrower sense in which the agent does not count as intending all of the goals she has settled on pursuing. Rather, she intends only a subset of her goals: those she can settle on achieving if she intends them, or those that rule out the pursuit of any incompatible alternatives. On the narrower usage, we may wish to deny that a young Sotomayor can intend to become a federal judge, since she knows that this outcome is not sufficiently within her control. We can bring this narrower sense out by contrasting 'intending' with 'trying', since it can sound more accurate in some contexts to say that Sotomayor was merely trying to earn a place on the bench. This is because she was not in a position to settle the matter on her own, and because (let's imagine) she did not positively foreclose the pursuit of other, incompatible careers. Her aspiration was a "goal state" of hers, but not a "plan state."

I am not convinced that there is a philosophically important ambiguity here, so I will try to remain neutral on this question in what follows. Keep in mind, though, that some of the debates about what intentions are might be trading on an equivocation between intentions as goal states and intentions as plan states.

3. Reductive accounts of intention

To accept that intentions are real attitudes is not yet to admit that they are a distinctive *kind* of attitude. It is still possible to hold that when we talk about intentions, we are really just talking about a variety of desire, or a kind of belief, or some combination of the two.

To understand what kind of attitude an intention is, we can make progress by thinking about its "direction of fit" with the world. Some attitudes, which we can call the "cognitive" attitudes, have a "mind to world" direction of fit: they represent their contents as true, or as philosophers often say, as "being the case." There are different ways in which we can represent some content – "I am in the French Riviera," say – as being the case. You might believe that you are in the French Riviera, or assume that you are for the sake of argument, or imagine that you are there. Each of these attitudes treats it as a putative fact that you are in the French Riviera, though for different purposes. In contrast, "conative" attitudes have a "world to mind" direction of fit: they represent some state of affairs as to-be-made-true, or to-be-realized. You might desire to be in the French Riviera, or wish that you were, or hope that you will be. These attitudes do not treat the idea of being in the Riviera as putative fact, but rather as something to be brought about.

When the notion of "direction of fit" is introduced, it is often said in the same breath that cognitive attitudes aim to match themselves to the facts, while conative attitudes aim to make the facts match them. But this claim is far too broad. Some cognitive states, like fantasies and assumptions, do not aim to match the facts as they really are; they represent their contents as fact for other purposes. Even though you know that you are not really in the French Riviera, there is nothing incorrect about imagining that you are. Likewise, not all conative states aim to bring the world into correspondence with them. The idle wish that snow were purple imposes no pressure

to try to bring such a change about. But the kernel of truth in this slogan is that it does apply to the paradigmatic cases of belief and desire.

Though G.E.M. Anscombe herself did not subscribe to these ideas as such, we can appropriate an ingenious example of hers to illustrate the point. Imagine a person shopping with a list of what to buy, and a detective who follows the shopper around while keeping a list of what he buys. The shopper's list aims to bring the world into conformity with itself, in the sense that the shopping cart contains all the items on the list. If the list says 'Raisin Bran' and the shopper puts Cheerios in the cart, the mistake lies with the cart and not the list. The remedy is to take the Cheerios out and put the Raisin Bran in, not to scratch out 'Raisin Bran' and write 'Cheerios' on the list. This is analogous to desire: a desire, if successful, attains its object. With the detective's list, it is the other way around. Her list aims to conform itself to whatever is in the cart, such that if the shopper puts Cheerios in his cart and the detective writes 'Raisin Bran' on her list, the mistake lies with the list and not the cart. The remedy in that case is to change the list, not to sneak up and switch the boxes in the cart. This is analogous with belief, as we ordinarily think of it: a belief, if correct, corresponds to what really is the case.

With these distinctions in mind, we can ask: are intentions more like the cognitive attitudes, more like the conative attitudes, both, or neither? And when they fail to correspond to reality, where does the fault lie?

a. Predominant desire

Intentions clearly have a world-to-mind direction of fit. They do not represent their contents as already being the case, but rather as to-be done. To be sure, they are also constrained by reality; for instance, it is a mistake of rationality to intend to do something if you fully believe that you cannot or will not do it. But desires are also constrained by reality, since there is something odd or defective in desiring what we know is already the case ("Oh, I want so badly for it to rain," she said, while staring out the window at the rain.). Further, when we fail to do as we intend, the failure seems to lie in what happened rather than in the attitude itself. The remedy for failing to carry out one's intention to go to the gym is to try again tomorrow, not to revise one's intention in favor of sitting on the couch. Thus, intentions bear many similarities to desire. This has led some to think, in a broadly Humean spirit, that we can simply identify intention with predominant desire. The desire in question must be predominant because, thank goodness, we do not intend to do everything that we desire to do. Perhaps, though, an agent intends to A just in case she desires to A more than she desires to do anything incompatible with A-ing.

The problem with this simple reduction, however, is that it fails to capture the element of *commitment* that is constitutive of intention. It is possible, even common, to predominantly desire to do something you do not intend to do, or even that you intend not to do. Think of the nicotine addict who has decided to quit, and so intends not to smoke another cigarette even though that is what he overwhelmingly desires to do. Further, we are able to form intentions even when faced with options that we desire equally. We are not paralyzed by menus that contain more than one

entrée that we would equally enjoy! Thus, intending to act seems to be distinct from merely desiring some object or state of affairs.

b. Predominant desire plus belief

Forming an intention disposes one to cease deliberating about what to do: one has thereby settled on an answer. To capture the sense in which to intend is to be committed or settled in a way that does not follow from mere desire, a natural idea is that we must combine desiring with believing. That is, perhaps having the intention to A is a matter not only of desiring to A, but also of believing that therefore, one will A. If Eveline is really settled on going to the party tonight, she should expect to be there – after all, it is up to her to ensure that she goes. And we do often express our intentions in a way that sounds like a prediction: we say "Sure, I'll go to the party" or even "I'll be there."

The idea that intending requires believing you will do as you intend helps to rule out cases in which the agent desires to do something but has not really made up her mind, and so continues to view it as an open question. If Eveline is asked whether she will go to the party and she replies "I don't know," her uncertainty implies that she has not yet formed an intention. On the other hand, it is not clear that it helps to deal with the case of the addicted smoker mentioned above. He not only most desires to smoke, but may well believe that he will smoke simply on the basis of statistics about recidivism and his own past history. Still, it seems possible that he is committed to quitting and so does not intend to smoke.

Further, some philosophers have offered examples that seem to illustrate the possibility of intending to do something without believing one will do it. Bratman points out that a given agent might be aware of her own forgetfulness, which leads her not infrequently to forget to carry out her intentions. Eveline might intend to stop and fill up her car with gas on the way home from work tonight but consider it a serious possibility that she will forget and drive straight home. This kind of case suggests that believing one will A is not necessary for intending to A. In a different kind of case, the agent might believe that she will make an attempt to A but lack the belief that she will succeed. Davidson's well-known example concerns a clerk who intends to make ten carbon copies by pressing forcefully with his pen on a stack of carbon paper (see also Chapter Six, section 2). Not having extensive experience with carbon-copying, however, he might be uncertain about whether his efforts will actually result in ten copies. Here, too, we seem to have a case of intending to A without believing one will A.

The latter kind of case might be dealt with by claiming that the agent intends only to *try* to A, and thus need only believe that she will try. After all, this is usually what we say when we wish to express doubt about our future success: "Well, I'm going to try to make ten copies, at least!" But while this is no doubt a useful turn of speech, it is not clear what it means to suppose that 'trying to A' is genuinely what the agent intends to do, if this is not simply the same thing as intending to A. Does the clerk who intends only to try to make ten copies behave any differently than the clerk who intends to actually make them? Assuming that the clerk is trying to make ten copies because he wants to actually end up with ten copies, then he will proceed by pressing as

hard as he can with his pen on the stack of paper, just as he would if he intended to do it. And if he does not end up with ten copies, and we ask him whether he has done what he intended to do, he will surely say "No" – even though he did try. Of course, we could imagine that he is only trying to make a show of it without caring whether his efforts succeed, and will thus put in minimal effort – but this is to make it a different kind of case.

The point is that if the difference between 'intending to try' and 'intending to do' does not show up in the agent's behavior other than to signal uncertainty about success, this suggests that it is not really a difference in the content of the intention. Rather, saying that one "intends to try" may simply be a mechanism for communicating doubt about success. Of course, there *is* a significant difference between the agnostic clerk and the confident clerk: a difference in their beliefs about what will happen, and the associated effects of those beliefs. The challenge is to make good on the claim that this difference in their beliefs amounts to a difference in what they intend to do.

c. Evaluative judgment

A third approach identifies the mental state of intention with a different type of belief: the belief that some available option is best or most choiceworthy. That is, to intend to perform action A is to judge that A is better than any other alternative to A-ing that one is considering. This is the view that Davidson embraced after becoming convinced that intentions are real attitudes. Though this view reduces intention to a kind of belief, it does not entail that the agent who intends to A must believe that she will A, and so escapes the problems raised for the "predominant desire plus belief" view.

There are several attractive features of this approach. One is that it fits well with a conception of agency as rational activity. According to one influential philosophical tradition, desires are not inherently rational; they are sometimes responsive to reason, but many times not. Indeed, this tradition conceives of desire as being in the category of "passion" rather than reason. If intending were primarily a matter of desiring, then this facet of our agency would be driven by whatever happens to attract us, rational or not, and it is unclear that this is significantly different from mere passivity. On the "evaluative judgment" conception, in contrast, agency is essentially guided by reason. A rational agent pursues whatever she takes to be the best thing to do, regardless of whether it is what she most wants.

Second, the view secures a natural fit between intention and practical reasoning, if practical reasoning is to be understood on the model of ordinary theoretical reasoning. When it comes to theoretical reasoning, we normally think of it as an activity of forming and revising our beliefs in light of the logical relations between other propositions that we believe. For example, if I believe the proposition "If the glove fits the defendant, then he committed the murder," and I also believe that the glove fits, then I might reason to the conclusion "The defendant committed the murder." If we think of intentions as evaluative beliefs, then we can understand reasoning about what to do on the same model: "If the glove doesn't fit, the best thing to do is acquit. The glove doesn't fit, so the best thing to do is acquit." In contrast, if intentions are a kind of desire, it is unclear how they could feature in reasoning of this kind; ordinary propositional logic does not tell us what

follows from the fact that something is desired, or how a new desire could be the conclusion of this kind of reasoning.

An apparent challenge for the view is that the belief "A is the best thing to do," where A is meant to occur at some future time, seems to be recklessly unjustified. After all, no concrete action A has occurred yet, such that the belief could be about *it*. Therefore, 'A' must stand for a type of thing that can be done – "going for a run," in general, rather than any particular run that has all the details specified. But it would be insane to believe that all possible instances of going for a run tomorrow would be better than the alternative. If it's 120 degrees Fahrenheit outside, or if someone is in an accident and urgently needs your help, or if you break your leg, or … then you should clearly not go for a run tomorrow. At best, then, we would seem to be justified only in drawing a conditional conclusion: "If conditions C obtain tomorrow, then going for a run is best." Rather than concluding that all our intentions are conditional, however, Davidson argues that they are *conditioned* on our background beliefs. Given that you do not expect any of the defeating circumstances mentioned above to arise, the set of possible runs your judgment refers to does not include instances that would be life-threatening or immoral.

A more insurmountable problem for this view is that it seems to be extensionally inadequate – it fails to capture all the cases in which we intuitively count as having an intention. For one thing, as we will discuss further in Chapter Nine, it seems possible to intend to do something even if you believe that some other available action would be better. A self-loathing drunk might intend to stop by the liquor store tonight to stock up while sincerely believing that it would be better not to drink. Second, even a fully rational and strong-willed agent will run into problems on this view. We frequently encounter situations in which we have multiple options that we believe to be equally good, or good in different, incomparable respects. A famous example of the first situation is handed down to us from Jean Buridan, who asked us to imagine a rational ass who is equally distant from two equally attractive piles of hay. If the ass must judge one of them to be better than the other in order to form the intention to eat from one of them, he'll starve to death. And a famous example of the second situation comes from Jean-Paul Sartre, who depicts a young man facing a choice between staying home to care for his sick mother and leaving to join the French Resistance during World War I. Plausibly, he knows that neither choice is better than the other, since the values they instantiate are incomparably different. Still, he can rationally form the intention to do one or the other.

These types of examples, which abound, suggest that intending does not require believing the intended action to be better than the relevant alternatives. We might try to deal with these cases by reinterpreting the view to hold that the intended action must be *as good as* any of the alternatives. But then the problem is that there is nothing to prevent us from intending *all* of the options in these cases, since each is at least as good as the others. This will land Buridan's ass and Sartre's young man in the same problem they had at the start, since intending both options will likely lead to accomplishing neither. What the cases show is that assigning value to our options can only get us so far, since we will face situations in which this approach delivers no unique answer to the question of what to do.

4. Plan states and plan rationality

Let us take stock. We have been attempting to understand what kind of attitude intention might be by reflecting on its similarities and differences with desire and belief. A different approach aims first to identify the functional role that intentions play, and then to ask what kind of attitude is suited to play that role. This methodology is exemplified by Bratman's "Planning Theory" of intention, which emphasizes the centrality for agents like ourselves of planning for the future. As noted earlier, we are creatures who frequently form intentions in advance rather than simply deciding what to do when the time comes to act. Why do we bother to do this? There would be no point in forming intentions in advance if they had no effect on what we actually do at the later time. Therefore, assuming that it does have a point, it must be that intentions serve as commitments in ways that help to control future action.

For instance, we might plan now for what to do later because we anticipate that our decision-making ability will be compromised at the later time. A student who waits to decide what classes to take until the moment comes to enroll will likely find himself making terrible choices, since he will not have enough time to research and think through the options carefully at the last minute. He will do much better if he deliberates beforehand and forms a plan about what to enroll in well before the time arrives. In a different kind of example, there might be enough time to decide later, but the agent may anticipate that she will be fatigued, intoxicated, subject to peer pressure, or otherwise likely to come to the wrong conclusion. If you know you are going to a party tonight, it can be beneficial to settle this morning on what beverages to consume and when to head home.

A second advantage of advance planning is that it facilitates coordination. Agents like us pursue many different goals at once – both for ourselves, and together with other people – and we will be much more effective at achieving all of them if we coordinate with ourselves and others over time. If we make no attempt to think through our future actions and bring them into harmony with one another, we will be prone to undermining our own efforts by acting in ways that are inconsistent. For example, I might have the goal of finishing a book on the philosophy of action within the next year while also being committed to spending quality time with friends and family. If I make no plans with an eye to balancing these two aims and decide what to do only at the spur of the moment, I will likely end up with no book or no friends when the year is over. Further, we are social agents that have need of coordinating with one another. If we are ever to have meetings together, play on teams together, teach classes or take them, we must make plans with each other.

Finally, planning needs to help us to structure our lives in situations like the ones faced by Buridan's ass and Sartre's young man. The kind of commitment involved must be possible even in the absence of a judgment that the option you are committing to is best, while ruling out the permissibility of intending options that are equally good but redundant or incompatible. The insight here is that we need not believe that there is a uniquely correct answer about what to do in order to make plans, nor do we need to think that everyone whose choices differ from our own is mistaken.

According to the Planning Theory, intentions are mental states that play these roles by serving as building-blocks in our plans. Thus, they are more than mere goal states. The hypothesis is that we can get a grip on the nature of intention by reflecting on what kinds of characteristics it must have, or ought to have, in order to play these roles well. A central claim of the planning theory is that intention facilitates these forms of intra- and inter-personal coordination in part by imposing structure on practical deliberation. In other words, forming an intention changes how we ought to go on in our reasoning about what to do. As noted earlier, forming an intention about what to do normally settles the question. Zhalisa might be torn between getting Mexican or Thai for lunch, but once she forms the intention to get Thai food, the question is now settled (absent some relevant new information, such as that the Thai place is closed today). If she maintains the intention to get Thai food while knowingly heading to the Mexican place, she has done something quite peculiar – especially if she had announced her intention to a friend and invited him to meet her at the Thai place.

The Planning Theory attempts to capture this insight by appeal to the idea that intending is subject to certain rational norms. 'Rational norm' here means a rule or requirement, the violation of which entails that the violator is in that respect irrational. The specific criticism of Zhalisa, then, is that she has been irrational in her planning and behavior. It is important to note that in making this criticism, we are not necessarily saying that Zhalisa had *most reason* to go to the Thai place, or that she ought all-things-considered to have gone there. It could be that the food there is atrocious, and her "friend" is toxic company she would be better off without, such that overall we think she did the right thing in going elsewhere. Simply forming the intention to do something cannot magically make it a good thing to do! We can nevertheless describe her as irrational in order to convey that her plans and behavior do not make sense from her own point of view. And the conjecture is that we can make progress on understanding what intentions are by more precisely articulating the rational norms that apply to them and not to other mental states.

To see what is distinctive about intention rationality, it will once again be useful to contrast intention with desire. It is common for creatures like us to have a variety of desires that are not mutually compatible. We want to be fit, and we also want to eat delicious sweets. We want to be wealthy and successful, but we want to spend our days having fun rather than working. We want to have the freedom of an unattached life, but the stability and comfort of committed relationships. While being torn between conflicting desires is often the source of both tragedy and comedy, there is nothing intrinsically irrational about it. We do not view people in this position as having made some kind of mistake that needs correcting; rather, it is a condition that we are inevitably subject to.

In contrast, having intentions that are not mutually compatible does seem irrational. Wanting to be in shape and wanting to sit on the couch all day is understandable, but intending to get in shape and also intending to do nothing but sit on the couch is criticizable. "You can't do both of those things!" we might say to such a person. "You'll have to choose!" We might put this thought by saying that there is a rational requirement that an agent's intentions be logically consistent with one another, given the agent's beliefs. Intuitively, a rational agent engaged in

planning will believe that all of her plans taken together are at least possible to realize (though see section 6 below). Call this the requirement of Intention Consistency.

Similarly, we are under no rational pressure to try to satisfy every desire that we find ourselves with. Simply experiencing the urge to pop all of the balloons at the party does not in any way obligate you to go into the kitchen and get an icepick. In contrast, we are under rational pressure to intend the means believed necessary to our intended ends. An agent who intends to pop all of the balloons in the next few minutes, but has no intention of going to get some sharp object, has plans that are incoherent. Even if we do not agree that he should pop all of the balloons, we might say "Your plan makes no sense! If you really do have the goal of popping the balloons, you're going to have to go get a popping implement." Call this the requirement of Means-End Coherence: intend the means that are believed necessary for accomplishing your intended ends.

Third, to play their role as building-blocks in plans, intentions must have a kind of default stability that desires need not have. There need be nothing amiss if a desire to write a novel arises only for a moment and disappears again, but intentions that stick around only for a moment cannot do their job of enabling coordination with oneself and others over time. In order to settle the question of what to do, they must be somewhat resistant to reconsideration. Having spent the time and cognitive resources deciding to attend the concert tonight rather than spending an evening at home, it would be inefficient to immediately abandon the intention and continue deliberating. It would be even more self-undermining to abandon the intention if you have already made further plans on the basis of the intention to go to the concert: bought tickets, told others that you would be there, and so forth. Of course, it is good to reconsider one's plans under some circumstances – when you acquire significant new information, or realize that your previous deliberation was likely to be mistaken. But on this approach, we violate a "Stability" requirement of practical rationality when we reconsider or abandon our intentions too easily, without good reason to do so.

The Consistency, Means-End Coherence, and Stability norms are formulated in a vague way here to leave room for argument about exactly how they ought to be stated. Note that in making these claims, we need not be saying that violating one of these norms makes one all-things-considered irrational. It might be that in some cases, there are good practical reasons to have inconsistent, means-end incoherent, or unstable intentions; occasionally, this might be the best way to achieve our goals. We could characterize this kind of situation as one in which it is all-things-considered rational to violate the requirements of rationality on intention, in the same way that it might be all-things-considered rational in some circumstances to maintain inconsistent beliefs – it might help you come across to your opponents in negotiation as an unpredictable maniac, for example. Still, the thought is that there is still some breakdown going on in these cases, and that this helps us to understand the kind of attitude intention is: namely, an attitude that is guided by the distinctive norms of plan rationality.

5. Cognitivism about intention

The Planning Theory articulates an important role that intentions play for us, emphasizing the sense in which they are a kind of rational commitment. In this respect, they are actually more

akin to belief than desire. After all, we could understand the mental state of belief as a commitment to treating a proposition as true. Further, belief is also a state that is subject to rational requirements. Our beliefs are rationally required to be logically consistent with one another, for instance; it is always irrational to believe contradictory things that could not all be true together. Some philosophers have been inspired by this observation to conclude that the role of intention could be played by a special kind of belief. Call this general claim 'cognitivism about intention'. This approach is similar to the view considered in section 3b. But whereas that view conceived of intentions primarily as desires with a further belief tacked on, the Cognitivist conceives of them primarily as beliefs or a belief-like state.

This strategy traces back to Gilbert Harman (1976), and more recent advocates include J. David Velleman and Kieran Setiya. Different Cognitivist views vary over the details, but as I am using the label here, what is central to all such views is the claim that intentions are attitudes that represent their contents as true: to intend to A is to represent it as true that one will A. It is in this sense that they are held to be belief-like. At first glance, this proposal is subject to the same objections as the reductionist view of section 3b, on which to intend is to believe that one's desire to A will lead to one's A-ing. However, this incarnation of the claim is bolstered by additional explanatory ambitions which, if successful, make it more difficult to reject.

One such ambition is to explain why intentions are subject to the aforementioned rational requirements of Consistency, Means-End Coherence, and Stability. Why is it irrational to plan in ways that violate these norms, rather than just zany and creative? The Cognitivist hypothesis is that if intentions are belief-like states, then the rational requirements on intention can be explained as deriving from the rational requirements on belief. For instance, perhaps our intentions are required to be consistent with one another because they are a kind of belief, and our beliefs are required to be consistent with one another. Nearly everyone will grant that it is irrational to believe inconsistent things, such as that it is currently raining here and that it is not currently raining here. If intending to fly to Shiraz tonight involves believing I will be in be Shiraz tonight, and intending to fly to Lahore involves believing I will be in Lahore and not Shiraz tonight, then my intentions are irrational because they involve inconsistent beliefs. Indeed, if I do not believe that I will end up in either place to which I intend to go, there would seem to be no problem with making plans in both cities for tonight. But this is generally not a good way to organize one's life and resources. Cognitivism offers a neat explanation for why good planning seems to demand that we rule out all the alternatives that are inconsistent with our plans.

Similarly, the requirement of Means-End Coherence might be explained as a prohibition on having explanatory gaps in one's beliefs. If I am committed to it being true that I will fly to Shiraz tonight, but have no beliefs about how this will come to pass – I am not committed to it being true that I will buy a ticket before the plane leaves, for instance – then my vision of the future does not hang together. If there is rational pressure on us to make our cognitive representations explanatorily coherent, this might explain why we should intend the means that we believe to be necessary for our ends. Otherwise, how could it be rational for us to expect our ends to come true?

It is trickier to account for the norm of Stability on intention by analogy to belief, since refusing to reconsider one's beliefs once they have been formed sounds like a kind of close-minded

dogmatism rather than a virtue. Arguably, our beliefs *should* change whenever our evidence changes (and not just when we have received significant new information bearing on what we ought to do). The best versions of Cognitivism grant this disanalogy and hold that intention-beliefs are distinctive in that they are not formed in response to evidence. Rather, they are formed and maintained on the basis of practical reasons: considerations that favor acting in a particular way. We represent it as true that we will go to the party tonight, not because we antecedently take ourselves to have sufficient evidence that we will go; after all, in that case, we would have no need of an intention to go. Rather, we represent it as true because we take ourselves to have sufficient reason *to* go. In this way, the Cognitivist can at least explain why the stability of our plans need not be hostage to the stability of our predictive evidence.

The promise of Cognitivism is that we might be able to trade two philosophical puzzles for one. Why are intentions and beliefs subject to somewhat similar rational norms? If it turns out that the answer to the former question derives from the answer to the latter question, we would seem to have made some philosophical progress. That said, we have only made progress if the explanation is adequate to the phenomena. Bratman argues that it is not, in that an agent can satisfy all the relevant rational requirements on her beliefs without satisfying the rational requirements on her intentions. If this is right, it shows that the requirements are not identical.

Begin with the idea that we can have false beliefs about our own mental states, including our own intentions (whether or not intentions are in fact beliefs). Just as you can believe you are in love when you are really just in lust, you can believe you intend to do something without really intending to do it. Now suppose that Michael truly intends to get in shape this month, which according to the Cognitivist means that he believes that he will get in shape this month. He also believes that it will be necessary for him to go to the gym every day in order to achieve this goal. Thus, in response to the pressure to be explanatorily coherent, he forms the belief *that he intends* to go to the gym every day and thus that he will go – where this is the 'will' of mere prediction. However, it might be the case that he doesn't actually intend to go; he hates working out, and even though he thinks he is committed to changing his ways today, he isn't. The problem for the Cognitivist is that Michael's beliefs are explanatorily coherent and consistent, and so in compliance with those rational requirements. However, he is still means-end incoherent: he has failed to actually intend the means he believes necessary for accomplishing his intended end.

Cognitivism about intention is also motivated by a second ambition, which is to account for the phenomenon of practical knowledge. As we saw in Chapter Two, some philosophers of action take as a starting point that intentional action involves a special kind of knowledge: if we are acting intentionally, we know what we are doing without the use of observation. If I have to look and see what I'm bringing about, this is at least a strong indication that it is not something I am doing on purpose. The need to account for this epistemic dimension of agency is a second major motivation behind the cognitivist approach. The undeniably elegant strategy is to point out that beliefs are the kind of state that can normally constitute knowledge. Thus, if intentions are a kind of belief, then intentions are also states that can embody knowledge. In particular, suppose that to intend to do A is in part to believe one will do A, and to intend to be doing A is in part to believe that one is doing A. If all intentional actions involve an intention so to act (more on this

question below), then all intentional actions involve a belief about what one is doing. And in the right circumstances, at least, this belief will amount to practical knowledge.

All of Chapter Six will be devoted to the debate over practical knowledge, so we will not delve into those details here. As we will see, though the Cognitivist explanation of practical knowledge is elegant, there are reasons to reject it. And in the end, the view is still subject to the kinds of counterexamples we examined in section 3b. It seems possible, and even sensible, to stop short of believing that one will perform each and every action one intends to perform. Of course, the Cognitivist may have further resources available to meet this objection. Setiya, for instance, proposes that intentions can take the form of partial beliefs as well as all-out beliefs, thereby allowing for a dubious agent merely to be more confident that she will do as she intends than she would be if she did not so intend. Still, we might worry that the parsimony of the Cognitivist approach comes at the cost of fully accounting for the complex ways in which a person's intentions and beliefs can differ from one another.

6. A distinctively practical attitude

If we reject all attempts to assimilate intention to other kinds of states or attitudes, the alternative is to countenance intention as a distinctive attitude in its own right. To indicate the ways in which it differs from desire and belief, we can characterize it as a "practical" attitude: its identity derives from its link to practical reasoning and action. The Planning Theory of intention is one such view, but there may be other ways of fleshing out the insight that intentions are *sui generis* practical attitudes. This kind of approach can grant that intention is often accompanied by various beliefs about what we are doing or will do, and even constrained by our beliefs. For instance, it is very plausible that we can or should only intend to perform action A if we believe it is possible that we will A. On the other hand, the Distinctive Practical Attitude view of intention maintains that intending to A is compatible with being agnostic or harboring doubts about whether we will succeed. The idea is that intention is a kind of commitment to acting that goes beyond mere desire, but that such commitment does not require blindness to one's own limitations and thus to the possibility of failure.

One advantage of this approach to intention is that it can make sense of the value of contingency planning – that is, planning for failure. It can be perfectly reasonable to sign a prenuptial agreement with your future spouse, and to pay the legal fees for doing so, even if you wholeheartedly intend to stay married. It is more difficult for the Cognitivist to explain why it would make sense to pay a price to protect yourself against failure, since to plan is to exclude failure on that view. On the other hand, from the perspective of parsimony, adding intentions to our model of human psychology in addition to belief and desire makes for more complexity. But this may be necessary to account for the complexity of human thought and action.

The Distinctive Practical Attitude view might also grant Setiya's point that intentions can come in degrees while denying that the degree to which an agent intends to A corresponds entirely her expectations about whether she will successfully A. We do not normally talk of partial intention, but there are several considerations that might push us in this direction. One such

thought is that we can be more or less committed, and if intention is a kind of commitment, it seems to follow that we can intend more or less. Perhaps it is not so unnatural to say that I more fully intend to feed my daughter tonight than I intend to work on this book tonight. Second, the Consistency norm on intention generates a problem that is analogous to the so-called "Preface Paradox" in the case of belief. The Preface Paradox concerns the fact that a person might be rational to believe each of a number claims – all those claims asserted in the person's book, for instance – while also rational to believe that she has made an error somewhere and at least one of those claims is false (hence, writing in the preface to the book that "the errors here are my sole responsibility"). This seems to contradict the rational requirement that our beliefs be logically consistent with one another, since the thinker is rational to believe both that each of her claims is true and that not all of her claims are true. Analogously, an agent might be rational in making each of a number of plans for the future while also rational in believing that she will not successfully execute all of these plans, thereby contradicting the Consistency requirement on intention. The move to partial intentions may help us to avoid this paradox, insofar as there is nothing amiss with strongly intending each plan while only weakly intending the whole package of plans.

7. Intending and intentional action

To conclude this chapter, let us take a step back and reflect on the relationship between intentions and intentional actions. We briefly considered the idea that these two phenomena are not really distinct from one another, but most of the views we examined take intentions to be attitudes rather than actions. Assuming they are distinct, how are intentions and intentional actions related to one another?

Clearly, intending to A does not entail that one will ever intentionally A, or even intentionally try to A. Sometimes our intentions lead nowhere: we may change our minds, forget, or procrastinate indefinitely. On the other hand, a very natural thought is that whenever one does intentionally A, one must have intended to A. Less schematically, the fact that Christine intends to finish her novel one day does not mean that she will, but if she does intentionally finish her novel, she must have intended to do just that. Following Bratman, we can call this latter claim the "Simple View" about the relationship between intentional action and intention (simple only in its straightforwardness, not its naïveté). The Simple View holds that for any description under which an action is intentional, the agent must intend to do that thing under that very description.

The Simple View is highly intuitive, and many other views in the philosophy of action presuppose it (for instance, as we shall see in Chapter Six, section 3a), the Cognitivist explanation of practical knowledge seems to depend on it). At the same time, there are reasons to doubt there will be such a simple one-to-one correspondence between intentional action and intention. On one way of looking at it, the two concepts are shaped by different interests and normative pressures. We have seen reasons for thinking of 'intention' as a broadly functional, psychological concept that is heavily defined by its role in coordinating action over time, both individually and together with others. The concept of intentional action, on the other hand, is connected to a variety of practical interests. For example, Joshua Knobe has collected strong empirical evidence that people tend to

use the concept in a way that is sensitive to their moral evaluation of the candidate action (the so-called 'Knobe Effect'). This is not to suggest that philosophy must adhere strictly in its theorizing to the folk concept of intentional action – we can certainly regiment the concept if we wish. But we should try to avoid departing so far from the ordinary concept that we end up theorizing about a notion we entirely invented.

The point is that even though the concepts of intentional action and intentions are closely related, we might expect this relationship to be less than perfectly simple insofar as diverse considerations pull them in different directions. More concretely, there are two main kinds of counterexamples that have been used to try to show that the Simple View is false. In one kind of example, the agent is in a situation where it is rational for her to aim at and pursue multiple goals at once, even though she knows these goals to be mutually incompatible. Suppose Yongxian is such a talented scientist that she can publish her articles in whichever journal she likes. Imagine further that she wants her newest article to be published as soon as possible, and there are two journals that are both sufficiently prestigious but highly unpredictable in the length of the review process. These journals coordinate their acceptances in such a way that while each review process is completely independent, any article that is accepted at one journal is immediately rejected by the other. Yongxian intentionally submits her article to both journals with the idea that this will result in the fastest possible acceptance.

There is nothing irrational about submitting to both journals, or about intending to submit to both – Yongxian's plan makes perfect sense. Further, when her article is published by *Nature*, we are inclined to say that her publishing there was intentional. After all, she intentionally submitted her article for publication there, and she is by hypothesis so talented that her articles are nearly always accepted. However, it would violate the requirement of Intention Consistency if she intended *to publish in Nature*, while at the same time intending *to publish in Science*, knowing that it is impossible to do both (notice that this is not the same thing as intending to *publish in either* Nature *or* Science, which is fine). Thus, if Yongxian is rational, she did not specifically intend to publish in *Nature* (or in *Science)*. If it works, we have a case of intentionally publishing in *Nature* without intending to publish in *Nature*.

A second kind of case concerns the foreseen but unintended side effects of our intended actions. We are often in a position to anticipate that our intended actions will have further effects or consequences that are not themselves intended. When these consequences are undesirable, we can face a difficult choice about whether to proceed with our plan and knowingly bring those consequences about. Suppose that Bas intends to uncork a bottle of champagne, foreseeing that this will make a loud noise and wake up the sleeping baby. Waking the baby is no part of his intention; he regrets this outcome and is not disposed to direct his efforts toward it in any way. If the baby unexpectedly did not wake up after Bas uncorked the champagne, he would not be disposed to uncork a second bottle or poke the baby. Still, if he wants the champagne badly enough that he decides to proceed and does thereby wake up the baby, we would be inclined to say that he intentionally awakened her (in fact, this is specifically what Knobe's research suggests in cases where the side effect is judged to be bad). After all, as his spouse might point out, he deliberately chose to do something that he knew would have this effect. However, by hypothesis, awakening

the baby is not what he intended. This appears to be a second kind of counterexample to the Simple View.

Those who reject the Simple View owe a different account of the relationship between intention and intentional action; the conclusion cannot be that what we intentionally do has nothing to do with what we intend. One possibility is that what we intend constrains and shapes what we can count as doing intentionally without requiring a one-to-one correspondence between the two. For example, it might be that Yongxian's action of publishing in *Nature* is only intentional because it is appropriately connected to her more general intention of publishing in one of the top science journals. If she had no such intention, and her article was submitted unbeknownst to her by a devious graduate student, then she did not publish in *Nature* intentionally. On the other hand, it might be that there are a number of distinct intentional actions that could all have been motivated and rationalized by Yongxian's general intention to publish her article. In general, if the Simple View is false, we must replace it with a more complex account of these dynamics.

Summary

A major divide in thinking about the nature of intention turns on whether intentions are psychological states. On one philosophical approach, the notion of intention can be reduced to, or understood in terms of, intentional action. This view tends to be associated with a methodology that prioritizes the present, taking as its central case intentional action that is happening now. In contrast, approaches that view intention as a real psychological state on par with belief and desire tend to prioritize thinking about future-directed intention. It is difficult to see how to account for the case of "pure intending" for the future in terms of intentional action when there is no candidate action now in progress.

Assuming that intentions are mental states, the next question is whether they are a type of belief or judgment, a type of desire, some combination of the two, or a distinctive kind of attitude in their own right. The problem with viewing intention as a form of desire is that intentions involve a kind of commitment that desires do not – they settle the question of what to do, and potentially render the agent criticizable if she does nothing to realize the intention. Further, it seems possible to intend to do something that one does not desire most to do. The problem with viewing intentions as a kind of belief is that beliefs ought to be based on sufficient evidence that they are true, whereas it only makes sense to form an intention when it is not already settled what will happen. Further, it seems possible to intend to do something that one does not believe one will do. The view that intentions are a combination of desire and belief inherits many of the problems with each of the first two views. Finally, the view that intentions are distinctive, irreducible attitudes is unparsimonious, but may nevertheless be justified by the drawbacks of the reductive approaches.

According to the most prominent way of thinking about intentions as distinctive attitudes, their identity is given by their role in future planning. We do not generally form intentions in isolation; rather, we construct a variety of plans, both for ourselves and together with others. In order to play their role as building-blocks in plans, aiding in intra- and interpersonal coordination

both at a time and over time, our intentions need to conform to certain rational requirements. They ought to be consistent with one another, given our beliefs. They ought to be means-end coherent, in that we ought not to intend an end without also intending any means we believe to be necessary for accomplishing that end. And they ought to be stable over time, in the absence of good reason to abandon them.

Although this kind of approach denies that we can understand intention in terms of intentional action, it is not necessarily committed to the reverse claim – that intentional action can be understood in terms of intention, or even that it must always involve intention. According to the Simple View, if an action is intentional under some description A, then the action must be the execution of an intention to A. This view has the advantage of offering a straightforward account of the relationship between intending and intentional action. However, there are counterexamples suggesting that this relationship is not in fact as straightforward as the Simple View would have it.

Suggested Reading

For defenses of the view that intending is simply a kind of doing, see Michael Thompson's *Life and Action,* Luca Ferrero's paper "Intending, Acting, and Doing," and Richard Moran and Martin Stone's "Anscombe on the Expression of Intention." Michael Ridge argues for the idea that intention is predominant desire in "Humean Intentions," while the view that intentions are a combination of desire and belief can be found in Wayne Davis's "A Causal Theory of Intending," Robert Audi's "Intending," and Neil Sinhababu's *Humean Nature*. Donald Davidson argues that intentions are evaluative judgments in "Intending," while Michael Bratman develops the Planning Theory of Intention in *Intention, Plans, and Practical Reason*, *Faces of Intention*, *Structures of Agency*, and *Planning, Time, and Self-Governance.*

The view that I have called Cognitivism about intention traces back to Gilbert Harman's paper "Practical Reasoning" and is given a forceful defense in J. David Velleman's *Practical Reflection, The Possibility of Practical Reason,* and "What Good is a Will?" See also Kieran Setiya, *Reasons without Rationalism,* and Berislav Marûsić and John Schwenkler, "Intending is Believing"

Bratman argues against the Simple View of the relationship between intending and intentional action in "Two Faces of Intention," while Hugh McCann offers a counterargument in "Settled Objectives and Rational Constraints." Joshua Knobe's experimental work on the folk concept of intentional action is also important to consider: "Intentional Action and Side Effects in Ordinary Language," "Intentional Action in Folk Psychology: An Experimental Investigation," and "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology." And Sam Shpall's "The Calendar Paradox" offers an interesting argument for thinking that intentions can be partial.

6. Practical Knowledge

Suppose you are intentionally walking to the beach. You happen across a friend who asks "What are you up to?" To answer her, you don't need to look down to see one flip-flop landing in front of the other, or observe the route you've traversed so far in order to deduce where you might be headed. Rather, you seem to be able to answer her without the use of observation or any kind of obvious inference. Moreover, your assertion that you are walking to the beach has a kind of default epistemic security. It would be odd for your friend to dispute your claim about where you are headed, or to ask 'How do you know?' absent some very unusual circumstances. Call this special kind of knowledge we have of our own intentional actions "practical knowledge."

Practical knowledge seems to be one of the most salient features of acting intentionally, at least under some descriptions. It is rare, or perhaps even impossible, to identify a case in which the agent is acting intentionally without any idea of what she is doing. This fact cries out for explanation. Indeed, as we saw in Chapter Two, some philosophers of action take practical knowledge to be the key to understanding the nature of agency, or at least a very promising starting point. This chapter will first attempt to clarify what practical knowledge is, and then examine various accounts of why practical knowledge is characteristic of intentional agency.

I. What do we mean by "practical knowledge?"

Practical knowledge can be contrasted with the more familiar phenomenon of "theoretical" or "contemplative" knowledge. To get our quarry into view, it will help to think about the features that both phenomena must have in order to count as species of the generic 'knowledge.' We can then think about the special features in virtue of which some bit of knowledge might count as practical rather than theoretical.

Answering the first question is tricky, since we should wish to leave open the possibility that practical knowledge is radically different from theoretical knowledge. Still, it is relatively uncontroversial to say that there are two features anything counting as knowledge must have. First, knowledge is factive: whatever is known must in fact be the case. One cannot know that P if P is false. No matter how good Misha's reasons are to believe that his uncle lives in Leningrad, he cannot know that his uncle lives in Leningrad if the name of the city has officially been changed to St. Petersburg, or if his uncle has recently relocated to Vladivostok. Second, knowledge cannot be the product of luck. Misha does not know that his uncle now lives in Vladivostok, even if this is true, if he arrives at that conclusion through mere guesswork or by asking a completely uninformed passerby. For knowledge, we must add that Misha has good justification for thinking it is true, or that there is a reliable connection between whatever Misha thinks is true and what is in fact the case, or that Misha arrived at this conclusion in a way that could not easily have led to a false belief.

Let us now turn to the features that distinguish practical knowledge from theoretical knowledge. As usual, a variety of proposals have been made here and there is far from universal agreement about all of them. Contemporary discussions of practical knowledge have all been

indelibly shaped by G.E.M. Anscombe's seminal remarks in her monograph *Intention*, so I will follow her in laying out a series of characteristics that practical knowledge has been said to have.

a. Knowledge without observation

The most prevalent characterization, and arguably the most fascinating, is that knowledge of intentional action is had "without observation." To illustrate what she has in mind, Anscombe offers the example of the knowledge we have of the position of our own limbs. Normally, she claims, there is no sensation or "kinaesthetic appearance" by means of which we find out where our limbs are. Likewise, she seems to be suggesting that practical knowledge is not obtained by way of sensory experience.

As we can see from the analogy, the class of things that can be known without observation is not limited to our intentional actions. It also includes those things that can be known *a priori*, such as mathematical facts. And arguably, it includes our knowledge of our own attitudes, such as what we believe and desire (though some think that we do know our own minds through observation).

It is natural to suppose from the example that Anscombe takes this claim to apply only to our intentional actions described as bodily movements, but this is not the case. Other examples she offers include the knowledge that one is painting the wall yellow, opening the window, and writing the word 'fool' on the blackboard with one's eyes closed. Indeed, she denies that we always have knowledge of how our bodies are moving when we act intentionally; sometimes, the most basic description under which we have non-observational knowledge will be at some remove from our bodies. You might know only that you are shooting a free throw, while having no non-observational knowledge of exactly how you are bending your knees and pushing with your arm.

b. Knowledge without inference

It is often claimed in the same breath that practical knowledge is not only non-observational, but non-inferential. That is, not only is it not based on observational evidence, but it is not based on anything at all. It is interesting to note that this claim features explicitly only once in *Intention*, in the following passage:

> When knowledge or opinion are present concerning what is the case, and what can happen – say Z – if one does certain things, say ABC, then it is possible to have the intention of doing Z in doing ABC; and if the case is one of knowledge or if the opinion is correct, then doing or causing Z is an intentional action, and it is not by observation that one knows one is doing Z; or in so far as one is observing, *inferring* etc. that Z is actually taking place, one's knowledge is not the knowledge that a man has of his intentional actions (50) (emphasis mine).

The view that practical knowledge is groundless is not given any real argument in the book, and it is difficult to see how to establish this claim solely by appeal to intuition. True, it does not normally *seem* to us as though we are explicitly running through an inference in our heads whenever we go to act intentionally. But explicit, conscious thought is not necessary for performing an inference; we make them all the time without awareness of doing so when we understand language, interpret subtle social cues, and so forth. Thus, although there is precedent for holding that practical knowledge must be non-inferential, this claim is on less solid ground than the claim that it is non-observational.

c. Mistakes are in the performance, not the judgment

Ordinary theoretical knowledge counts as such in virtue of being in accordance with the facts it represents. If you know that electrons have a negative charge, this is in virtue of your correctly tracking reality. To employ a distinction introduced in Chapter Five, theoretical knowledge has a mind-to-world direction of fit. In contrast, Anscombe claims that when there is a discrepancy between what the agent believes she is doing and what is actually happening, the mistake is in the performance and not the judgment. One has failed to do the correct thing, not to have believed the correct thing. In our terms (not Anscombe's), practical knowledge has a world-to-mind direction of fit.

This classification seems innocuous enough, but at times Anscombe seems to be making an even more radical claim. In discussing the example of writing the word 'fool' on the blackboard with one's eyes closed, she claims that the agent's knowledge that she is doing so would be the same even if something goes wrong with the chalk, such that the words do not appear. This remark goes significantly further than saying that the mistake is in the writing; it suggests that you can still have practical knowledge even if you are completely failing to do what you think you are doing! If this is the right way to think about practical knowledge, then it is difficult to see how it could be a species of the generic 'knowledge'. Again, as we ordinarily think about knowledge, you cannot know something that is not true.

d. The cause of what it understands

Anscombe approvingly quotes St. Thomas Aquinas, who characterized practical knowledge as "the cause of what it understands." This is contrasted with 'speculative' knowledge, which is "derived from the objects known." Aquinas discusses the idea of practical knowledge in the context of accounting for God's knowledge of the world He created. If God could only come to know of the world by receiving information about the objects in it, this would make Him dependent on things outside of himself. But God could not be dependent in this way. Aquinas argues instead that God can know how things are in virtue of having created them. Since there is no question of the world failing to be as God has willed it, He need only consult the ideas in His mind in order to have knowledge of what he has made. Anscombe seems to suggest that although we humans are certainly not as infallible as God in our production, we too can have knowledge of

what we accomplish in virtue of being the ones who accomplish it. That is, we know what we are doing intentionally because we are making it happen.

The notion of 'cause' here is potentially ambiguous. On one hand, we might mean "efficient cause," or as Aristotle defined it, "the primary source of change or rest" – that which sets things in motion. God was certainly understood to be the cause of the world in this sense. But we might also mean "formal cause," that which unifies some entity into a single thing and gives it its form or "shape." God was also thought to be the formal cause of the objects in the world – that in virtue of which giraffes and trees have the form that they do. It is a matter of interpretation whether Anscombe had both of these senses in mind, or only one, in speaking of practical knowledge as the cause of the action known.

e. Contradicted by interference

When a person expresses an ordinary theoretical belief, like "The Whitney is located at the corner of 75th and Madison," one can contradict the assertion with a competing assertion of fact: "No it isn't, it moved to Gansevoort Street in 2015." In contrast, Anscombe claims that the agent's description of her own intentional action, like "I am making a soufflé for dinner" is not properly contradicted by the assertion of a competing fact, such as "No you're not, for I just saw it collapse in the oven." Rather, the contradiction is the expression of a competing intention, such as "No you won't, for I won't let that happen – I'm going to cut the power to the oven!" Would-be expressions of practical knowledge, in other words, are properly contradicted by another would-be expression of practical knowledge.

2. The scope and object of practical knowledge

With the phenomenon of practical knowledge somewhat in view, we can now ask: what exactly is it that an agent knows practically when she acts?

A very tempting first thought is that we have special, non-observational knowledge only of what we *intend* to be doing, or what we are *trying* to do – in other words, something that is in our own minds. It is not surprising that we do not have to observe what our bodies are bringing about in order to know what action we have in mind. In contrast, it is much more difficult to understand how we could know without observation whether our interventions have succeeded in bringing the world to be as we envisioned it. One possible conception of practical knowledge, then, is commonly called the "two-factor view:" that the practical, non-observational component is limited to what the agent has in mind, while any further knowledge of what actually happens must be obtained in the ordinary perceptual or inferential way. Alternatively, a modified version of the two-factor view might allow the practical knowledge component to include the movements of the body.

Anscombe strongly rejects both versions of the two-factor view, calling it a "mad account." However, her reason for this seems to derive from skepticism about the very idea of intending or willing as mental items that are separable from acting. Those who are less skeptical about the

reality of these psychological notions might therefore find the two-factor view more attractive than Anscombe did.

She herself seems to be of the view that practical knowledge includes "what happens." This need not mean that we are infallible when we take ourselves to know what we are doing; things do sometimes go awry. It is possible to think you are comforting your friend by saying certain things, only to have exactly the opposite effect. But it is equally the case that theoretical knowledge need not be infallible, on pain of radical skepticism. Still, when things go well and one does in fact intentionally effect some change in the world, Anscombe holds that one knows what one does, and one does what happens – there is no distinction in that case between your comforting a friend and the friend being comforted. That said, this claim is primarily meant to apply to the more immediate descriptions under which we act intentionally rather than to the far-out effects that we are ultimately aiming at. Anscombe grants that we commonly fail to achieve our more distal goals, such as making partner at the law firm or bringing about a political revolution. However, she holds that it is "necessarily the rare exception" for our more immediate actions to fail to be what we suppose them to be.

A second question concerns just how intimately practical knowledge is connected with acting intentionally. A number of philosophers have held, following what seems to be the implication of Anscombe's remarks, that having practical knowledge is a *necessary* condition of acting intentionally. That is, its absence would suffice to show that an occurrence was not an intentional act. If an agent is asked 'Why are you pouring salt all over the floor?" and she replies "Oh, I didn't know that I was," this is sufficient to determine that she was not pouring the salt intentionally. Among those who hold that it is a necessary condition, there is room for disagreement about exactly what the necessary condition is. The strongest version of the claim is that for any act-description A, A-ing is only an intentional action if the agent has practical knowledge of A-ing. Others are willing to grant that we can fail to have knowledge of our intentional actions under some descriptions – even those that are relatively immediate – but claim that we must have practical knowledge under at least one description. And a still weaker view has it that we might lack knowledge under all descriptions, but holds that there must be at least one description under which we *believe* without observation or inference that we are so acting, or have a higher credence than we otherwise would. Finally, the skeptic denies all of these claims, holding that it is possible to act intentionally without practical knowledge, belief, or increased credence under any description.

To sort through this thicket of possibilities, it will be helpful to work through some concrete examples. First, given its prevalence in this debate, it is obligatory to discuss Donald Davidson's example of the carbon copier (which first cropped up in Chapter Five, section 3b). The example concerned a clerk who wants to produce ten carbon copies. He stacks up ten sheets of carbon paper under a piece of regular paper and presses hard with his pen on the top sheet with the aim of making ten copies of what he is writing. However, he is uncertain whether it will work to make ten copies at once, and so does not know or even believe as he is writing that he is in fact making ten copies. He will only come to know whether he is by flipping through the sheets and observing the results of his efforts. But if he does end up making ten copies, in just the way he was aiming to

do, it is plausible that he has done so intentionally. If it works, this kind of example shows that one can intentionally do A without having practical knowledge or belief that one is A-ing.

For those who wish to maintain that having practical knowledge is necessary for acting intentionally, one possible response is to deny that the clerk actually makes ten copies intentionally. We might think that his doubt precludes him from doing it intentionally, since it shows that he lacks the requisite knowledge of *how* to make ten copies at once. He is guessing that it is possible to do it in the way he is attempting, but he does not know, and so is not sufficiently in control of the outcome to count as intentionally bringing it about. A second kind of response is to insist that he does know that he is making ten copies. He lacks knowledge that he is making ten copies *at once*, but as long as he has the opportunity and the will to check the results and keep at it until he ends up with ten copies, then perhaps he does know all along that he is making ten copies. Either of these interpretations are strategies for defending the strong version of the claim that practical knowledge of A-ing is necessary for intentionally A-ing.

Alternatively, one might give up on the strong claim and grant that the copier intentionally makes ten copies without practical knowledge that he is doing so. There does seem to be something he knows that he's doing, however – pressing hard with his pen on the stack of paper, for instance. This suggest that a weaker necessary condition might be true: if an agent is intentionally A-ing, there must be some (perhaps very low-level) description of her action – Z-ing, say—such that she has practical knowledge that she is Z-ing. That is, if there is no description of what she is doing under which she knows or believes that she is so acting, then it cannot be an intentional action of hers.

This claim is fairly plausible when it comes to Davidson's carbon copier, but there are other possible counterexamples that call it into doubt. One kind of counterexample concerns highly skilled, "zoned-in" action, such as that of the expert athlete or craftsperson. When Serena Williams once again hits the ball to just the right spot on the tennis court, we would generally grant that she does so intentionally; after all, she is fully in control of the ball and hits it to just the right spot because it is the right spot. Yet skilled athletes often claim that they are unaware of exactly what moves they are making or why, at the time of making them – hence the need to watch tapes of their performances. Of course, they know what they are doing under a very broad and abstract description like "playing tennis." But that does not help the opponent's case, since the knowledge that one is playing tennis does not depend upon the occurrence of any particular result, as the knowledge that one is making ten copies does. In fact, one interpretation of what is going on in such cases is that top-down, concept-laden awareness of action would interfere with the athlete's ability to perform. Rather than being necessary for intentional agency, practical knowledge would actually interfere with it in these circumstances.

A second kind of example concerns the "zoned-out" agent who is acting habitually, or as we might say, "on autopilot." The commuter who has driven home from work countless times may be entirely unaware of what she is doing as she takes each prescribed turn on her route, uses her turn signal correctly, responds to the flow of traffic, and so forth. If asked why she is making a particular turn, she must first observe where she is and which turn she is making in order to answer the question. Nevertheless, her driving is controlled, purposive, and responsive to her reasons, and

so she may aptly be described as driving that route intentionally (though admittedly, if 'being intentional' is a property that comes in degrees, this example might be best understood as less than *fully* intentional).

A third kind of counterexample is more outré, but attempts to show directly that doubt can penetrate all the way down to the most basic act-description available. Imagine an agent who has the disorder called 'Alien Hand Syndrome', which causes him to experience his hand as performing goal-directed actions without his control. One of the acts his hand commonly performs is buttoning up his shirt, which happens independently of his wanting or choosing to button it. On this occasion, he does want his shirt to be buttoned up, and he is confident that this will happen if he does nothing. However, he prefers to button it up under his own control. If he does try to button his shirt with his anarchic hand, he succeeds a fair amount of the time, but less often than if he simply lets the hand act on its own. Suppose that he does succeed in buttoning his shirt with his hand under his control, just as he had in mind, and so does it intentionally. Yet as he is doing so, he is *less* confident that he is moving his hand in such a way as to button his shirt than he would be if he were not acting at all. Since this is the most basic description under which his action is intentional, there is nothing he is doing intentionally that he knows, believes, or even has an increase in confidence that he is doing.

If counterexamples like this work, then we should be skeptical of the claim that practical knowledge under at least one description is necessary for acting intentionally. At most, we should grant that we normally have practical knowledge of what we are doing when we act intentionally. This concession is important because it opens up more possibilities for explaining how agents come by practical knowledge. For instance, a necessary connection between acting intentionally and acting knowingly cannot be explained by appeal to a merely causal mechanism, since causal mechanisms are in principle subject to breakdowns that would fail to preserve the necessity of the connection. But if we have independent reason to think that breakdowns in that connection are possible, then a causal explanation is on the table after all.

3. Accounts of practical knowledge

Why is it the case that we always or normally have practical knowledge of our intentional actions, under at least some description? This section will look at three views about the answer to this question, each of which locates the explanation in a different place. The first kind of view explains practical knowledge by appeal to the nature of intention; the second explains it by appeal to the nature of events that are in progress; and the third explains it by appeal to an evidential connection between intention and intentional action.

a. Cognitivism about intention

As we saw in Chapter Five, there is disagreement over how to understand what intentions are. One view, which we called 'Cognitivism about intention', holds that intentions are a kind of belief or importantly akin to belief. Roughly, Cognitivism holds that intending to A just is, or

essentially involves, being in a belief-like state that one will A. One major attraction of this view is that it is in a good position to explain the phenomenon of practical knowledge.

The key is that according to Cognitivism, intention-beliefs are the kind of state that can constitute knowledge. Suppose the Simple View is true (see Chapter Five, section 7), such that intentionally A-ing always involves the intention to A. That is, if Matty is intentionally disrupting the faculty meeting, it must be the case that he had the intention to disrupt the meeting. It follows, according to Cognitivism, that intentionally A-ing involves believing that one will A: forming the intention to disrupt the meeting now involves forming the belief "I'm going to disrupt the meeting now." This belief is not based on observation of the intended action, and we are generally happy to think of intentions as the cause (either efficient or formal) of intentional action. Thus, Cognitivism seems to offer an elegant explanation of how practical knowledge could be both non-observational and the cause of what it understands.

As noted in Chapter Five, there are independent reasons to doubt that Cognitivism is the best way to understand what intentions are. Most importantly, it seems possible to intend to A without believing that you will A – you might think that there is a substantial risk of failing to A, or of forgetting to even try. But even if we set such objections aside, Cognitivism faces a challenge in explaining the epistemology of practical knowledge: why should we expect that intention-beliefs will ordinarily amount to *knowledge*?

As we noted at the beginning of this chapter, merely having a true belief that P is not enough to know that P. Knowledge requires that the thinker be connected to the truth in a way that goes beyond a lucky guess. One must have good evidence that P, or good reason to think that P is true, or one's belief must be the product of a method that reliably produces true beliefs, or it must be the case that one could not easily have been wrong about whether P. The question for the Cognitivist is whether intention-beliefs will normally meet some such justification or safety condition.

The first thing to note is that intentions are not based on prior evidence for their truth. We normally form the intention to A because we desire to A, or because we believe that we have sufficient reason to A. You intend to travel to Istanbul because you want to see the Blue Mosque, or because you believe that you ought to visit your grandmother. In fact, if you already have sufficient evidence that you will go to Istanbul – perhaps you know that your grandmother plans to have you kidnapped and brought there – then you have no need to form the intention to go. Thus, if intentions are beliefs about what one is going to do, they are beliefs that are formed on the basis of a desire that they come true, in the face of insufficient evidence that they will come true. We ordinarily call this kind of belief-formation "wishful thinking" and consider it patently irrational rather than a way to acquire knowledge.

However, the Cognitivist can point out in response that these are no ordinary cases. Usually, having the belief that P doesn't bring it about that P is true. But intention-beliefs do play a causal role in bringing about the contents they represent. Having the intention to travel to Istanbul does contribute heavily to bringing it about that one in fact goes to Istanbul. Thus, as J. David Velleman has argued, there *is* a reliable connection between intending to A and A's being true – it simply runs in the opposite direction from the normal case. By acting out our intentions,

we make them come true. Perhaps this is enough for them to constitute knowledge. Setiya makes the potentially complementary suggestion that "knowledge-how" provides the epistemic warrant needed for an intention-belief to amount to knowledge. He argues that we are only entitled to form the belief that we will A if we know how to A, and that meeting this constraint exempts us from the general requirement that our beliefs be grounded in sufficient prior evidence for their truth.

Still, intending to do something is generally not sufficient for getting it done (would that it were!). The road to hell is paved with all kinds of good intentions that never came to fruition in the way the agent had in mind. And in many instances, the act-type in question is one we know quite well how to do. On some occasions, we do not even try to carry out an intention because we have forgotten about it. I might intend to call my Aunt Linda to wish her a happy birthday when I get home from work, only to realize on the following day that the call had completely slipped my mind. In other cases, when the time comes to act, we experience "weakness of the will" and cannot bring ourselves to do what we intend. A nervous student might form the intention to speak up on the first day of class in order to make a good impression on the professor, but find himself simply unable to raise his hand and speak when the time comes. In a third kind of case, we try to do what we intend but fail to bring it off, even if it is something we know how to do in general. It is one thing to intend to make a funny joke at a dinner party, or to stay married to someone for life, and another thing to succeed. In light of these various ways in which intentions can fail to be realized, we may worry that intentions cannot constitute practical knowledge even when they come true, since they are not reliably enough connected to succeeding. At the least, these kinds of cases suggest that the justificatory story must be more complicated than merely pointing to the self-fulfilling aspect of intending.

b. Imperfective knowledge

The Cognitivist approach faces this epistemic challenge because it aspires to make good on the claim that practical knowledge extends to "what happens" – to what we are going to do, and to what we will have done. An alternative approach, advocated by Michael Thompson, attempts to avoid some of these worries by restricting the content of what is known to descriptions couched in the imperfective aspect. What we know practically is not what we did, are going to do, or will have done, but what we *are doing*.

'Imperfective' is a grammatical term for constructions that characterize activities as incomplete and ongoing, or "in media res." It contrasts with the perfective aspect, which is used to characterize activities as a completed whole. It is the difference between 'I am going/will be going/was going to the protest' and 'I went /will have gone to the protest'. What is of interest about the imperfective in this context is that one can *be going* to a protest without ever having gone. This is because true imperfective act-descriptions do not entail the truth of a corresponding perfective description (this is sometimes referred to as the "imperfective paradox," though there is nothing especially paradoxical about it). If I am on my way to the protest when I get called away to an emergency surgery, or get hit by a bus, or merely change my mind about going, it is

nevertheless true that I was going up until that point. Moreover, true imperfective act-descriptions are compatible with fallow periods, in which one is not making any progress toward one's goal. I can be going to the protest even if I am standing still at a stoplight at the moment, or if I go a bit out of my way in order to chat with the crossing-guard.

Thompson proposes that practical knowledge concerns only what is presently going on, and thus exists only when the action in question is not complete. In fact, he maintains that "Its character as knowledge is not affected when the hydrogen bomb goes off and most of what the agent is doing never gets done" (2011, 209) – the hydrogen bomb being simply a rather violent way of referring to possible interruptions. The agent's knowledge is not affected by failure because the object of knowledge is an unfolding process that can be happening whether or not she succeeds in achieving the goal. And this means that in order to have practical knowledge, we need not be justified in supposing that we will actually pull it off. That said, the opportunity to try, try again if at first one does not succeed also plays a role in securing the claim that even when we are currently failing, we may yet prevail and thus count as having been doing the action all along.

We might still worry about certain cases of catastrophic failure, however. Suppose Luca sets out to deadlift a three-hundred-pound barbell, thinking that it can't be too difficult because he just saw his friend Gabe do it. As Luca steps up to the barbell, he confidently asserts that what he is doing is lifting three hundred pounds. As it turns out, Luca is far from capable of lifting that amount of weight; not only does the barbell refuse to budge, but there was never any real possibility that he would lift it on this occasion. It does not seem true in Luca's case that there was ever a time in which he was lifting the weight, or that there was a process of lifting that was ongoing but incomplete. After all, repeated attempts would not make it any more likely that he will succeed (indeed, it would be even less likely). Similarly, when Don Quixote charges at a windmill, taking himself to be fighting a giant, it rings false to say that he is in the process of fighting giants, no matter how many times he tries.

Cases in which success is impossible due to the agent's general inability or unfavorable circumstances seem to show that success cannot be completely irrelevant to practical knowledge. On pain of completely divorcing practical knowledge from material reality, Luca doesn't know that he is lifting the barbell and Don Quixote doesn't know that he is fighting giants. In response, Thompson seems to suggest that such cases fall outside the scope of Anscombe's use of 'intentional', and so can be legitimately excluded. He writes of Davidson's carbon-copy case that if the clerk only has one shot to make the ten copies, and isn't allowed to check and see if he's succeeding as he goes along, then this "will not be an illustration of the topic of Anscombe's book, any more than lottery-winning is when you bought the ticket with that aim" (2011, 210). In the kind of case that he takes to be connected with practical knowledge, the agent will have repeated opportunities to fail and try again.

An intriguing implication of this approach is that the relevant sense of intentional action turns out to be essentially self-conscious, in the way that many have thought our own minds are self-conscious. As Thompson puts it, attributing the definition to Anscombe, "What's up with me is an intentional action precisely where it is a content of specifically practical knowledge, otherwise not" (2011, 203). Just as we might think that it is essential to pain that the sufferer is aware of

feeling it, the idea is that there could be no such thing as intentional actions we are unaware of performing. Such self-consciousness is normally thought to be confined to what is going on in our heads, but the radical claim here is that it extends to what is happening in the external world.

c. The Inferential Account

By emphasizing the ways in which imperfective act-descriptions are relatively non-committal, the Thompsonian approach makes it easier to see how we could know what we are doing without observation or any other kind of evidence. One drawback is that it gives up on the plausible idea that we can have practical knowledge not only of what we are doing, but of what we will do in the future. In some circumstances, at least, it seems that I can know that I will be at the party tomorrow – and clearly, not by observing my own future. A second drawback is that it must restrict the notion of intentional action, arguably beyond its everyday usage, in order to exclude problematic cases. And a third is that it gives up on including the actual achievement of material results in the content of practical knowledge, which is part of what makes Anscombe's proposal so intriguing.

A third approach attempts to avoid these drawbacks by giving up on a different aspect of Anscombe's remarks: the idea that practical knowledge cannot be the product of inference. As noted above, the claim that the knowledge must be non-inferential is made only once and in passing in *Intention*, and it is far less supportable merely by appeal to intuition or introspection. After all, many cognitive processes are inferential without feeling as though an inference is being made in conscious thought. If you hear someone utter the sentence "I saw her duck," you must infer from the context who 'her' refers to and whether what the speaker saw was an aquatic fowl or an evasive movement. And yet this process normally feels perfectly immediate; we are often not even aware of considering interpretations other than the one we actually arrive at. Could it be that practical knowledge is similarly inferential?

An inferential account will be most appealing to those who independently subscribe to a non-cognitivist view of intention (see Chapter Five, sections 3a and 5). If intention is a kind of desire or a distinctively practical attitude, it is not the sort of doxastic state that could embody practical knowledge in itself. Intentions could, however, serve as part of the *evidential basis* for practical knowledge. If intending to A normally leads to one's intentionally A-ing in conducive circumstances, then we might come to know that we are A-ing intentionally, or will A intentionally, by inferring this conclusion from the premise that we intend to A.

To avoid some of the problems that plagued the first two accounts, the Inferential Theory denies that the premise 'I intend to A' suffices on its own to justify the conclusion 'I will A'. It will only be against a background of other relevant beliefs that we are justified in inferring from intention to action. For one thing, the agent must believe that she has the ability to do what she intends – or at least, she must lack good reason to doubt that she has the ability. The clerk in the carbon-copy example doubts that he is able to make ten copies in one go, and so rationally refrains from inferring that he is actually doing it. Second, in order to know that one is doing what one intends, one must believe that the circumstances are conducive to success – or at least, lack good

reason to doubt that they are. Even if the carbon-copier is fully confident that he can make ten copies at a time, he should not believe he will make them if he thinks he might well be prevented by a faulty pen or his arch-nemesis. A third relevant factor is the agent's history of irresoluteness in the relevant context. An agent who plans to quit smoking should not believe that she will quit if she has a long history of intending to quit and weakly failing to do so.

On the other hand, the Inferential Theory can grant that under less committal imperfective descriptions, knowing what one intends to be doing is frequently sufficient for knowing what one is doing. If setting out with the intention to walk to the beach suffices to be in the process of doing so, even if you are currently standing still or walking in the wrong direction, then knowing what you intend is enough to know you are in the process of walking to the beach. It is only under more committal descriptions that entail accomplishing something – a completed event – that these further conditions must hold in order to have practical knowledge. To have knowledge that you really are making ten copies, in the sense that entails ending up with ten copies, you must know that you intend to do it and are able to do it in the relevant circumstances.

In order to know about one's abilities and circumstances, experience is required. This might seem to suggest that according to the Inferential Theory, practical knowledge turns out to be observational after all. Perhaps we can know what we intend without observation, but to know what actually happens, we have to use our senses. This would be to misunderstand the proposal, however. Although experience of one's own abilities and observation of the circumstances are required in order to be in a position to make the relevant inference, it is not required that you observe the action itself. The view can vindicate the intuition that an expert carbon-copier (unlike Davidson's topic) can know that he is making ten copies, and will have made them, without looking at the bottom sheet.

Summary

Ultimately, the decision about which of these explanations of practical knowledge is most appealing will depend heavily on one's other philosophical commitments. In many ways, cognitivism about intention offers the most elegant account of practical knowledge, but there are independent reasons to doubt that an intention to A is or essentially involves a belief that one will do A. For those who independently subscribe to alternative views of intention on which intention is a kind of desire or a distinctively practical attitude, the Inferential Theory of practical knowledge will likely be a better fit. On the other hand, the question of whether having practical knowledge is necessary for acting intentionally is also highly relevant. The Inferential Theory is unlikely to satisfy those who think it is, since there is nothing in that approach to ensure that the inference will be made in each case of acting intentionally. The Thompsonian account may seem to do better on this score since it restricts the definition of intentional action to actions that are practically known. But if there really is a philosophical motivation for this definitional constraint, the Inferential Theory could also take it on board. The drawback is that such a constraint arguably renders the theory too narrow by departing from ordinary ways of thinking about intentional action. And though all accounts would do well to grant the insight that knowledge couched in the

imperfective is much easier to come by than perfective knowledge, we might also worry that imperfective knowledge does not go far enough – that we sometimes have practical knowledge not only of what we are doing, but of what we have done and will do.

A further question concerns the epistemology of practical knowledge, and whether it should be governed by standards that are similar to those governing theoretical knowledge. We should be open to the idea that practical knowledge is distinctive and should not merely be assimilated to the more dominant framework of theoretical knowledge. On the other hand, the two must share some similarities, such that it is legitimate to understand both of them as species of the genus 'knowledge'. The Inferential Theory effectively collapses practical knowledge into ordinary theoretical knowledge by requiring that it be evidentially justified in the ordinary way. The other two approaches depart more radically from a traditional epistemic framework, allowing that knowledge of the empirical world can be had without sufficient prior evidence or can even be completely groundless. This opens such views up to the question of why it should count as knowledge at all.

Suggested Reading

Obviously, one must begin with G.E.M. Anscombe's *Intention* on this topic. Stuart Hampshire's *Thought and Action* is also an important starting-point. The two-factor view is defended by Keith Donnellan and Sidney Morgenbesser in "Knowing What I am Doing." The Cognitivist solution is most fully developed by J. David Velleman, particularly in *Practical Reflection*, as well as by Kieran Setiya in *Reasons Without Rationalism*. Rae Langton offers an important critique of this approach in "Intention as Faith;" see also Sarah Paul, "Intention, Belief, and Wishful Thinking." Kevin Falvey's "Knowledge in Intention" and Michael Thompson's "Anscombe's Intention and Practical Knowledge" both emphasize the importance of knowledge that is couched in the imperfective aspect. H. P. Grice offers a version of the Inferential Theory in "Intention and Uncertainty, and Paul's "How We Know What We're Doing" defends a somewhat different version.

7. Does Action Have a Constitutive Aim?

Some kinds of activities and objects are defined, at least in part, by their *aim* or *purpose*. For example, perhaps you are reading this with a steaming mug of coffee near to hand. The purpose of a coffee mug is, roughly, to contain a hot beverage in such a way as to allow a human being to drink from it without burning her hand. There are a variety of materials and designs that are compatible with this purpose. Mugs can be made out of ceramics, steel, or plastic, and can take a number of different shapes. There is a limit to this variety, however; one cannot simply fashion a paper towel into a cylindrical shape and call it a mug, since this could not serve the purpose of containing a hot beverage. And some mugs serve their defining purpose better than others. Mugs that have no handles, and no alternative means of protecting a human hand, are clearly worse as mugs than those with handles. Correspondingly, the activity of making a mug is defined in part by the aim of producing an artifact that serves this purpose to some satisfactory degree.

In cases like this, we come to understand a great deal about what the kind of activity or object is by coming to understand its purpose. If a small child wants to know what a mug is, or what a mug-maker is doing, explaining what a mug is *for* would be a very good start. Moreover, understanding the aim of a mug tells us something about what it is to be a good mug or a bad one, and how to go about making a good one. In other words, we can derive norms of being a good mug, or being good at mug-making, from the purpose of the mug: for example, we can determine that mugs ought to have handles.

The central question of this chapter is whether action and agency have a defining aim or purpose in this sense. In the literature on this topic, the somewhat cumbersome phrase that is typically used is "constitutive aim," meaning "aim that (partly) constitutes the nature of the thing." If action has a constitutive aim, then investigating that aim will be deeply relevant to our grasp of what action is. And it will also be deeply relevant for the branch of philosophy that seeks to understand the foundations of ethics. Moral demands on us primarily concern our actions – in particular, how we act where the interests of others are concerned. Plausibly, to act well is in part to act morally, and to act immorally is in part to act badly. So if we can identify a constitutive aim of action that will tell us something about what it is to act well or badly, then perhaps we will have made progress on understanding what it is to act morally and why we should care about doing so.

1. The Guise of the Good

Aristotle begins the *Nicomachean Ethics* with the rather grand claim that "every action and pursuit is thought to aim at some good." This has come to be known as the thesis that action takes place "under the guise of the good," or "*sub specie boni*." According to this thesis, to be an agent is to be in the pursuit of goodness or value. To act, therefore, one must represent what one is doing as having some value. Of course, one might be wrong in any given case; an agent who sets out to accomplish some good by buying a certain stock, thereby making a profit, might be completely mistaken in thinking that the stock is a good investment. What is essential is that the action *appear*

to be good in some respect from the agent's point of view and be undertaken in light of this apparent goodness. The investor who intentionally buys a stock must believe that she will profit from it (or that some other good will come from owning it), even if she is wrong about that.

The strongest version of this thesis is that intentional action aims at what is judged to be the *best* option, or what is most worth doing. Plato's Socrates seems to have espoused this view, holding that "… no one who knows or believes there is something else better than what he is doing, something possible, will go on doing what he had been doing when he could be doing what is better" (*Protagoras* 358b7-c1). As we will discuss more thoroughly in Chapter Nine, this claim is deeply counterintuitive. Surely we do sometimes eat, drink, smoke, and blow off our homework when we know it would be better not to! But according to the Socratic thesis, when we intentionally do such things, we do not in fact know that it would be better not to. When we fail to do what is best, it is always out of ignorance or the inability to do otherwise.

A weaker version of the Guise of the Good thesis holds only that the agent must represent her action as good in some respect or other. This allows for the possibility of intentionally and knowingly doing what is not best as long as you think there is something good about what you actually do. You might think that overall it is better to read than to watch TV, since you will learn more from reading and knowledge is more valuable than pleasure. But you might nevertheless choose to watch TV in light of the fact that it is enjoyable and in that respect good.

Both versions of the thesis are motivated by a particular view about what desire is in rational creatures. As we saw in Chapter Three, our actions are not only prompted by our desires but also rationalized by them. When Ben gets on his bike and rides to his favorite Mexican food truck, he does so because he wants a burrito and believes that he can get a burrito at the food truck. Ben's desire for a burrito justifies his action, or makes sense of it, whereas it would not make sense if he had no desire for Mexican food. But we might ask: *how* do desires rationalize action? One natural thought is that they do so by portraying their object as good, thereby showing what good the agent thought he would accomplish by acting as he did. After all, rational creatures want what is good for them.

Again, Plato's Socrates seems to have thought that rational creatures desire only what is best for them. A weaker, more empirically plausible claim is that while it is possible to want things that we know not to be best, it is impossible to want something that is not deemed to be good in any respect. Recall an illustration of the latter claim that we first encountered in Chapter Three, in which G.E.M. Anscombe asks us to imagine someone who claims to want a saucer of mud – not for any further purpose, and not because it is taken to be pleasant, fitting, or morally required, but simply wanted. She claims that this is "fair nonsense." And if it is unintelligible to want what is not taken to be good in any respect, then it seems to be equally unintelligible to act in pursuit of an end that is taken to have no value.

To further understand the motivation behind the thesis, it may be helpful to consider an analogy with belief. It is often claimed that the attitude of belief has the constitutive aim of truth, such that a belief is correct only if it is true. If you believe that the Earth is flat, then your belief is wrong no matter how convinced you are that it is true and no matter how much evidence you think you have. The reason it is incorrect to believe the Earth is flat, the thought goes, is that to

believe that P is to try to get it right – to believe is to aim to believe truly. This view depicts the activity of believing as guided by a standard of truth that any believer must at least implicitly accept. It is by reference to this standard that we can understand believing as a rational activity and assess particular instances accordingly. Further, in virtue of accepting truth as the constitutive standard of believing, a thinker who is paying attention to a particular belief of hers must find it incoherent to think of her belief as false. Just try it: "I believe the Earth is flat, though that's not true – the Earth is not flat." This combination of thoughts cries out for the thinker to make an adjustment in what she believes. Likewise, the claim is that action is an essentially rational activity that is guided by the standard of goodness or value, and that must be viewed by the agent herself in this light.

Criticisms of the Guise of the Good thesis have taken two main forms. One prong of criticism takes the form of enumerating counterexamples to the claim that desires and intentional actions are undertaken in light of some perceived goodness. One class of counterexamples purports to show that the acknowledged good can fail to attract or motivate us. Michael Stocker offers the example of a politician who believes that helping other people is good, and who is in a position to help, but who no longer has any desire to do so. Perhaps he has become bitter and misanthropic in his advanced age, or perhaps he is depressed or apathetic, such that the plight of others now fails to move him. This kind of example poses more of a difficulty for the stronger Socratic thesis that the believed good *must* attract than it does for the weaker view that *only* the good attracts. In response, the Socratic might defend the thesis by arguing that the interference of emotion or mental illness can mask or block the normal response of a rational agent to the perceived good, essentially rendering him irrational in that respect.

A second class of counterexamples purports to show that we can be motivated by what we take to be bad, precisely because it is bad. A paradigmatic case might be Satan as depicted in the Christian tradition: a being who desires and pursues only evil. More worldly examples include self-destructive, masochistic, or sadistic desires to harm oneself or others. Anscombe – a proponent of the Guise of the Good thesis – considers the case of Satan and points out that in *Paradise Lost*, Satan proclaims "evil, be thou my good!" She concludes that Satan is no counterexample since for him, the good of what he does *is* that it is evil. But this response threatens either to trivialize the thesis or to do Satan an injustice. It trivializes the thesis if Satan gets to count as conforming to the thesis just in virtue of having any desires and pursuits. It would turn out that all we mean by that thesis is that 'the good' is simply 'whatever is desired or pursued', which is not very interesting from an ethical perspective. On the other hand, suppose Anscombe means to be claiming that Satan really does judge all of the things he desires and pursues to be good, in the same sense that God judges quite different things to be good. On this reading, as J. David Velleman points out, Satan turns out to be just another sap who is trying to do the right thing. Surely, Velleman suggests, more perversity is possible than this!

The second prong of criticism of the Guise of the Good thesis takes the more general form of articulating conditions that are sufficient for agency but that do not entail the thesis. Velleman carries out this strategy by arguing that action has a substantial constitutive aim other than goodness; we will discuss his view in the next section. Kieran Setiya is skeptical that action has a substantial aim, but argues that it is possible to act intentionally without believing that one has a

justifying reason for doing it. As we saw in Chapter Five, Setiya follows Anscombe in holding that to act intentionally, one must have practical knowledge (or at least belief, or increased credence) of what one is doing, and one must have an answer to the question 'Why are you doing that?' But to satisfy these conditions, he suggests that the agent need only have a causal-psychological *explanation* available for what he is doing, not a justification. One can answer the 'Why?' question by saying 'Because he killed my brother!' without believing that this constitutes any kind of justification for one's act of revenge. The agent might in fact be convinced that it doesn't, but nonetheless aware that he is moved to act by considerations of revenge.

This claim seems to conflict with Anscombe's insistence that we cannot make sense of desiring something, like a saucer of mud, without being able to cite some respect in which we judge it desirable. But Setiya finds fault with her example on the grounds that, as he claims, desires are necessarily for actions or outcomes and not simply for objects (though of course, you can desire to *have* an object). He counters that if we redescribe the case in this way – as a desire to have a saucer of mud in one's hand just for a moment, say – it is not so clearly unintelligible. As long as the agent can know that she has this desire and is acting in order to satisfy it, this is enough to be engaged in rational agency; she need not have the further belief that having a saucer of mud would be good in any way. It might well be true that human beings tend to pursue the good, but Setiya concludes that this is not essential to rational agency as such.

2. The aim of self-understanding

Velleman denies that action constitutively aims at the good, but he does think that action has an aim. In fact, he argues that it *must* have an aim. The case for this rests on what he takes to be the inadequacy of the Davidsonian model on which beliefs and desires combine to constitute reasons and thereby to produce actions. The problem, he argues, is that this story leaves the agent out of the picture (see also Chapter Four, section 6b). The agent's role is to adjudicate between the reasons she has and the actions she performs, resolving conflicts between various options and determining what she shall do. As Velleman puts it, "When I participate in an action, *I* must be adding something to the normal motivational influence of my desires, beliefs, and intentions …" (1992b, 465, emphasis mine). This does not mean that we must be committed to a picture on which there are such things as irreducible agents with mysterious causal powers. Rather, if we are after a reductive account, we must seek a state of the person that can play the functional role of the agent. And this state, he argues, is a desire or aim that helps to regulate the person's behavior.

Characterized abstractly, Velleman argues that the aim is to act in accordance with reasons (though not necessarily "goodness-tracking" reasons, as we shall see). Nothing could count as a rational agent that did not at all have the aim of doing what it has reason to do – for what would be rational about a creature that acted without any regard for reasons? Moreover, this aim is not something that we can rationally disavow or become alienated from. To rationally disavow the aim would involve taking oneself to have good reason to disavow it, and this would be to manifest a concern for reasons after all. On the other hand, we can manipulate ourselves into not caring

about our reasons through techniques like taking drugs, but this would be to suspend our agency altogether. This is the sense in which having the aim constitutes us as rational agents.

On Velleman's own view, the aim is actually more specific than this in human agents. What we aim for is not to act for reasons, conceived of as *reasons*; rather, we aim to be intelligible to ourselves. We want to know what we are doing, and what we will do in the future. Not only this, but we want to understand why we are doing it. We want to avoid being surprised by our own behavior or otherwise finding ourselves indecipherable. As evidence for this, he points out that when we find ourselves lacking awareness of what we are doing – think of going into the kitchen, only to be completely unable to remember what you are there for – we generally stop what we are doing until we can figure out what we are up to. Thus, the aim of finding ourselves intelligible gives us extra motivation to do what we think would make sense to do and to refrain from doing anything that would not make sense. It "plays the role of the agent" by throwing its weight behind those actions that would render us understandable to ourselves and against those actions that wouldn't. And when it does this, we are in fact acting out of a concern for reasons, although we might be thinking of the relevant considerations only as bearing on what "makes sense."

This may sound similar to the Guise of the Good view, but Velleman argues that it can make perfect sense to perform actions that you do not view as good in any way. If you know that you are a masochist, then you would find yourself highly intelligible if you engaged in self-harm even if you see no actual value in harming yourself. In general, there is nothing to guarantee that an agent's sincere beliefs about what is good or of value will coincide with her beliefs about what kind of person she is or what she is motivated by. It follows that one can be guided by considerations bearing on what it makes sense to do without thereby believing that the recommended actions are in fact good.

Though in earlier work, Velleman described the goal of self-intelligibility in terms of a desire the agent has, he later came to characterize it as a "drive." This is important because the first reaction of many to the view is that they don't *feel* as though they have a particularly strong desire for self-intelligibility, and certainly don't feel as though this desire partly motivates every action they perform. But a drive is not something that we would expect ourselves to be vividly aware of, or to think explicitly about when deciding what to do. Rather, it is meant to be understood as a motive, or goal-directed fund of psychic energy, that regulates cognitive functioning and behavior without requiring conscious awareness. Think of the way that we all vigorously regulate the distance between ourselves and the person we are speaking to. We generally do this without thinking about it at all, unless we encounter a person from another culture who has a different idea of how much personal space is appropriate. Our pursuit of self-understanding will generally also be backgrounded in this way.

Whereas the Guise of the Good view of agency seems to exclude too much (chiefly, perverse and weak-willed action), the main objection to the self-intelligibility view is that it threatens to include too much. Intuitively, there are lots of things that would make sense for us to do but that we have no good reason to do. If our agency is oriented toward avoiding confusion and surprise, then whenever we expect ourselves to behave in a certain way, we will be better agents to the extent that we actually do it. This would seem to include pessimistic predictions like "I'm definitely going

to screw up my class presentation!" as well as the continuation of vices that one otherwise wishes to get rid of – "I'd like to quit smoking, but since I'm an addict, it wouldn't make sense for me to quit." Correspondingly, we would seem to be functioning poorly as agents to the extent that we pursue substantial change in ourselves or our lifestyle. The smoker who does manage to take steps to quit will not count as acting at all, if the aim of self-intelligibility does not play a contributing role in motivating that behavior. These counterintuitive implications challenge the view to explain how exactly the aim of self-intelligibility can motivate action that seems to conflict with that goal.

3. The aim of self-constitution

Christine Korsgaard offers a third conception of what the constitutive aim of action could be. She argues that action essentially aims at constituting oneself as a unified person and integrated agent. In fact, she claims that it is impossible to act without this kind of unity, since action must be attributable to the person as a whole and not to some part of her. Action must have an author. But at the same time, we achieve this required unity by acting. Korsgaard takes her initial cue from Plato:

> "One who is just does not allow any part of himself to do the work of another part or allow the various classes within him to meddle with each other. He regulates well what is really his own and rules himself. He puts himself in order, is his own friend, and harmonizes the three parts of himself like three limiting notes in a musical scale – high, low, and middle. He binds together those parts and any others there may be in between, and from having many things he becomes entirely one, moderate and harmonious. Only then does he act." (Plato, *Republic* 443d-e).

Action succeeds in constituting oneself as a rational agent to the extent that it is chosen in a way that renders one both efficacious and autonomous. In acting, we determine *ourselves* (autonomy) to be the *cause* of some end (efficacy). To choose an action in the right way, we must conform to principles that tell us how to be autonomous and efficacious. And on Korsgaard's view, these principles turn out to be Immanuel Kant's Hypothetical and Categorical Imperatives.

The Hypothetical Imperative tells us how to be efficacious in achieving our ends: "Whoever wills the end also wills the indispensably necessary means to it that are within his power." This principle is similar to the Means-End Coherence principle we considered in Chapter Five, expressing the basic idea that to commit yourself to realizing some goal is also to commit yourself to taking the steps required in order to bring that goal about. If your end is to attend law school in the United States, you must take the LSAT, since this is a necessary means of being admitted to law school in the U.S. The Categorical Imperative, on the other hand, tells us how to be autonomous or self-determined: "Act only according to that maxim through which you can at the same time will that it become universal law." In choosing to act on some consideration – deciding to help another person because she is in need of aid, say – we must view that choice as legitimate for *any* rational creature to make in similar circumstances. Only then is it the case that we are

acting autonomously, since only then is our behavior attributable to us qua rational agent and not to internal or external forces acting upon us.

Those who are familiar with Kant's ethics will have noticed that the Categorical Imperative is better known as a moral principle. According to Kant, it is the fundamental moral principle: the only actions that are morally permissible are those that are chosen in conformity with, and out of respect for, the moral law embodied by the imperative. In addition to capturing what it is for oneself to act autonomously, the imperative tells us how to act in a way that manifests respect for pure reason and autonomy in others. Kant claims that the above formulation of the Categorical Imperative is equivalent to saying that we must act in such a way as to treat humanity, whether in your own person or the person of another, always at the same time as an end and never merely as a means. We must not use other autonomous beings as mere instruments for our own ends, by coercing, deceiving, or otherwise treating them in ways that they do not autonomously consent to.

Putting it all together, we get the somewhat surprising result that one cannot act without psychic unity, and one cannot have psychic unity without acting in conformity with, and out of respect for, the moral law. An immoral person is a disunified person, and at the limit, such a person is not capable of acting at all. Whatever she does fails to be done autonomously, and thus amounts only to heteronomous behavior (behavior that is attributable to a force external to the agent herself). This implication poses a challenge for Korsgaard's view, which she inherits from both Kant and Plato: if genuine action must be morally good, how is it possible to perform an evil act? Or if there is no such thing as evil action, how can we ever be blameworthy – which we surely are? Korsgaard attempts to meet this challenge by finding a middle ground between full autonomy and complete psychic anarchy. Following Plato, she compares the organization of a person's psyche to the constitutional order of a state. A state can have a constitution that allows it to function in a minimal way while exhibiting significant defects in its laws and organization. For example, a tyrannical government can unify a population into something that counts as a state while failing to constitute it as a well-functioning, harmonious, or effective polity. Likewise, a psyche can be unified enough to count as acting, but that unity might be the result of defective principles that result in defective action. Examples include a person who only ever pursues her own pleasure, or who always defers to the advice of others without using her own powers of reason. Thus, the idea is that action comes in degrees, such that an action that fails to constitute the person as autonomous and efficacious is both bad and defective qua action.

But, we might wonder, how can action constitute you as a unified agent if being a unified agent is necessary for acting in the first place? The thing to remember is that on Korsgaard's view, being a unified agent is not necessary for the phenomenon of moving one's body around in a purposive way and causing effects in the world. This is something even the most disunified wanton can do. The question, as she sees it, concerns the conditions under which the causing of effects in the world is attributable to the person as a whole – as the author of those movements and effects. And the answer is that these events are only attributable to an agent if they are chosen according to principles that render her autonomous and efficacious.

Finally, Korsgaard's view faces a challenge in accounting for the possibility of agency in young children and non-human animals. Like Velleman's, the view is highly intellectual, requiring

sophisticated reasoning and representational capacities. If we are willing to grant that non-human animals exhibit a form of agency, it cannot be because they are capable of guiding their choices with respect to the Hypothetical and Categorical Imperatives. Rather, some other account must be given of how animal action is possible without the capacity for reason. Her most recent work aims to do just this, arguing that intelligent instinct and teleologically-loaded perception enable non-rational animals to engage in a kind of action. Where rational creatures choose both their action and the purpose with which they act, in light of taking the purpose to constitute a good reason to act, non-rational animals can choose only the act component; their purposes are given to them by way of evolutionary forces and by their perceptions of the world. They are autonomous in the sense that they are governed by the principles of their own nature, but the content of those principles is given to them by instinct. We rational creatures, in contrast, can and must choose our own maxims, and this is a far deeper form of autonomy.

4. The will to power

We have seen that constitutive-aim views of action have roots in the work of Plato, Aristotle, and Kant. Paul Katsafanas has argued that yet another conception of the aim of action can be found in the work of Friederich Nietzsche. According to Katsafanas, Nietzsche held that all action manifests the "will to power," understood as seeking out and overcoming resistance in the pursuit of some end. He thought that human actions are not motivated only by the desire to achieve some outcome, such as having written a book. They are also motivated by psychological forces that Nietzsche called "drives," which aim at their own expression. Drives differ from desires in that desires tend to specify some state of affairs or end, the achievement of which would satisfy and temporarily extinguish the desire. Drives, in contrast, are directed at the performance of the activity itself and are thus never truly satisfied or extinguished. Many of our actions will involve both end-setting desires and drives. For example, the activity of writing is aimed at the end of finishing the book, but it may also be motivated by the drive for glory, leading the writer to proceed in a way that seeks out and overcomes obstacles to being glorified.

The Nietzsche-inspired claim is that all human action is drive-motivated, and that the process of pursuing our goals and prevailing against resistance is an essential aim that we have in addition to actually achieving our ends. This aim gives us reasons to seek out actions that afford the opportunity to confront and overcome resistance. These are not the only reasons we have on Katsafanas's view; the agent's other values provide independent standards of good and bad action. But when there is a conflict between these standards and the will to power, the latter must take priority in virtue of its inescapability.

5. No constitutive aim

Though many have been attracted to the idea that action has a constitutive aim, it is certainly not obligatory to think that it does. The conceptions of agency that do posit such an aim are relatively complex and sophisticated; they require that all possible agents have surprising

motivations at work behind the scenes. This tends to have the effect of reserving the status of 'agent' for creatures who are self-conscious and have advanced reasoning and cognitive capacities. Such a conclusion will be unacceptable to those who seek a more inclusive understanding of agency, on which simpler beings like non-human animals and artificial intelligences count as acting in broadly the same sense that adult human beings do. Those who embrace a "thin" conception of action, on which all it requires is something like belief-desire motivation or motivation by intention, will likely wish to deny that there *must* be any additional aspiration going on when a creature acts (though there might well be in any given case). We will see how some such views attempt to account for the difference between mere action and more complex forms of autonomy in Chapter Eight.

6. Implications for ethics and metaethics

In addition to offering insight into the nature of agency, views on which action has a constitutive aim are exciting to those who seek to understand where normative reasons for action come from and why we should be motivated by them. For instance, what makes it the case that we have reason not to do immoral things like murdering one another? According to a widespread conception of morality, moral demands are objective. What does it mean to say they are objective? On one reading, to be objective is to be "part of the fabric of the world," to use J.L. Mackie's memorable phrase, rather than being dependent on the human mind. The existence of subatomic particles is objective in this sense, whereas the existence of the US dollar is not. We generally do not think of moral requirements as being dependent on whatever we happen to believe, desire, or agree to; we think that people should not commit murder even if they really want to, and that communities (like the Mafia) that do not prohibit murder are wrong not to do so. That said, it can be difficult to see how our reasons to act in some ways and not in others could be found in the natural world, just like objective facts about subatomic particles and chemical reactions. Facts about how the world is generally do not entail anything on their own about what human beings ought to do. This is why many people have found supernatural explanations appealing: if there is an authoritative God who stands outside of nature, then perhaps the demands on us to act in certain ways come from such a being. Of course, this kind of supernatural explanation involves mysteries of its own.

On an alternative reading, the objectivity of morality does not require that it be part of the fabric of the world in this metaphysical sense. Rather, to say that such demands are objective is to say that they are inescapable for any creature to whom they apply, in the sense that all such creatures must occupy a practical point of view from which those demands are necessarily viewed as authoritative. In other words, one cannot avoid the moral requirement not to murder simply by failing to care about morality in the way that one can fail to care about the demands of etiquette or fashion. A rational agent will necessarily be motivated by the demands of morality; she cannot fully understand that it is morally wrong for her to murder and be completely unmoved by that fact. And if she asks *why* she must care about doing what is morally right, a successful answer will connect the property of moral rightness to something that she cannot help but care about, or be

moved by, or see as authoritative from her own practical perspective. Here, the challenge is to say what such an answer could possibly be.

Enter constitutive aim theories of action. The basic strategy here is to derive reasons for action from the aim of action by showing that the aim of action gives us standards by which to evaluate actions as better or worse. Think back to the example of the mug. A vessel is better at serving the aim of a mug if it has a handle, and this means that anyone who has the aim of making a mug has reason to add a handle. Analogously, if action constitutively aims at self-intelligibility, then anyone engaged in action has reason to do what would make sense. And if action aims at self-constitution, where this requires conformity with the Categorical Imperative, then anyone engaged in action has reason to act only on maxims that could at the same time be willed as universal law. In general, our reasons for action will be considerations that tell us how to meet the standards set by our aim in acting.

If it works, the strategy has hope of vindicating at least the second, ethical sense of objectivity and perhaps even the first "metaethical" one. With respect to the first sense, the claim is that our status as rational agents is both naturally and practically inescapable for us. Short of some drastic final act like suicide or self-lobotomization, we cannot help but be agents and thus to have the aim that is constitutive of being an agent. And if you ask "why should I be an agent who has this aim?" where what you are looking for are practical reasons to be an agent with that aim, you have thereby demonstrated that you already *have* the aim of doing things for reasons. Thus, the idea is that the standards that constitute us as rational agents are the things that we cannot help to care about, be moved by, or see as authoritative from our own practical perspective, and this is what gets us the objectivity of the reasons that we can derive from these standards.

Further, constitutivism might offer an appealing compromise concerning the first, metaphysical sense of objectivity. On this approach, human beings do create reasons for action through the activity of reasoning about what to do rather than simply discovering them in God or nature. Reasons therefore pose no special problem for naturalism. That said, our reasons are not simply dependent on what we happen to believe, desire, or agree to, and so do not end up being merely "subjective" either. They apply to all agents simply because we must act, and in acting, we must aim at some standard.

Though these possibilities are exciting, there are reasons to be skeptical that the project could succeed. One worry is that the appeal to inescapability could not on its own allay the worry that we are merely "victims of a collective 'practical' delusion, helplessly devoted to pursuing something which actually ought not to be pursued," as Matthew Silverstein puts it. Further work must be done to show that the reasons we are committed to are actually valid and not a massive error. A second worry is that we could not have genuinely normative reasons to conform to the constitutive standards of an enterprise – even an inescapable enterprise – unless we had antecedent reason to engage in that enterprise. David Enoch illustrates the objection by imagining beings called "shmagents" who do not care about being agents and are happy performing "shmactions," which are non-actional events that are as similar to actions as they can be without conforming to the constitutive standards thereof. The shmagents might ask why they should be agents who act instead, and if there is no genuinely normative reason that can be given in answer to that question,

then they seem to have no reason to conform to the constitutive standards of agency. Much of the ensuing discussion of "metaethical constitutivism" has centered around whether and how to respond to the shmagency objection.

Summary

In many if not all instances of action, there is some particular goal the agent is pursuing: she is acting in order to satisfy her hunger, or to arrive at her destination, or to learn the answer to a question. The thought we have explored in this chapter is that in addition, every possible action must be motivated or regulated by some more general goal. One longstanding view is that action necessarily aims at what is good. Whenever we act, in other words, we are trying to do achieve something that is genuinely of value. A second suggestion is that we are seeking self-understanding. On this view, we may not always try to do what is good or worthwhile, but we are always trying to do what makes sense given what we know about ourselves and our situation. A third suggestion is that we are striving for unity of self. To succeed in acting, we must at the same time impose the kind of structure on our various desires and inclinations that amounts to an autonomous self who is doing the acting. And a fourth proposal is that all action seeks in part to express our Nietzschean drives to overcome resistance and expand our power.

These ideas are exciting because they offer the hope of telling us what it is to be not just *an* agent, but an excellent agent. If all action has some essential purpose, then we act well just in case we fulfill that purpose completely. An excellent agent is one who does genuinely good things, or who is fully intelligible to herself, or who is fully unified around her capacity for reason. And if that's right, then we have made significant progress in understanding how we *ought* to act – since, plausibly, we ought to act well. That said, these implications for ethics and metaethics are only as convincing as the theories of action that generate them. We must be careful not to get interesting implications out only by building more into our account of agency than is actually necessary.

Suggested Reading

For further reading on the Guise of the Good thesis, the papers in Sergio Tenenbaum's collection *Desire, Practical Reason, and the Good* are an excellent place to start. Most of the papers in the collection argue in defense of the thesis, but see Kieran Setiya's "Sympathy for the Devil" for a critical perspective. J. David Velleman's "The Guise of the Good" and Michael Stocker's "Desiring the Bad" are also classic discussions that offer a critical take. Sections 36-41 of G.E.M. Anscombe's *Intention* are relevant as well.

Velleman's view of agency and its constitutive aim is developed over a series of books and articles, including *Practical Reflection*, the collection of papers in *The Possibility of Practical Reason*, and *How We Get Along*. For critical discussion, see Michael Bratman's "Cognitivism about Practical

Reason." Christine Korsgaard's view of rational agency is articulated in *Self-Constitution: Agency, Identity, and Integrity*, while the extension to non-human animals is defended in *Fellow Creatures*. More on Paul Katsafanas's view can be found in *Agency and the Foundation of Ethics: Nietzschean Constitutivism.*

For critiques of the strategy of deriving conclusions about reasons from the aim of action, see David Enoch's "Agency, Schmagency: Why Normativity Won't Come from What Is Constitutive of Action" and Matthew Silverstein's "Inescapability and Normativity."

8. Identification and Self-Governance

A recurring idea throughout previous chapters is that actions must be attributable to an agent or a person rather than to some external force or mere internal mechanism. In contrast with undergoing digestion or being buffeted by the wind, action is *self*-movement. Some ways of thinking about action suggest that this is an all-or-nothing matter – an occurrence either is attributable to the agent, and thus an action, or it isn't. Either some force is acting on you, pushing you around, or you are pushing yourself around. But we might also be interested in a kind of agency that comes in degrees. It is natural to think that some of our intentional or voluntary actions are more autonomous, or more expressive of our true selves, than others. Suppose that if it were up to Archana, she would be together with Colin. But because her family puts a great deal of pressure on her and threatens to disown her otherwise, she enters into a marriage with Amar. She does this voluntarily and out of a desire to appease her parents, but as we might say, unwillingly. This kind of action still counts as intentional, but there is something missing; it is not a case of exercising agency in a full-blooded way. We therefore need an account of the difference between such cases and the manifestation of agency *par excellence*.

This chapter will explore that idea using the notion of "self-governance," a phrase which explicitly invites a reading on which it comes in degrees. Think of a territory that was formerly a colony of another sovereign state, but that has now been granted some amount of independence. We can ask of such a territory to what extent it is self-governed, and to what extent its policies, laws, and actions are at odds with the will of its people or its leaders. Likewise, the idea of self-governance as applied to the individual leaves open the possibility that a genuine action might still be more or less reflective of the person's true will. Archana's actions are to some extent her own, but they are also governed by the wishes of her family. Under what conditions, then, are we fully autonomous?

There is a sense in which this was also the topic of Chapter Seven, which explored the idea that action has a constitutive aim. Such views tend to say that an agent is acting more autonomously to the extent that he is more successful in meeting the standards set by the aim of action, whatever that happens to be. This chapter takes up a similar question, but begins with skepticism – or at least agnosticism – about the idea that intentional action as such aims at anything in particular. How might we nevertheless capture the idea that some actions are more expressive of the agent's true will than others?

1. Frankfurt on identification

As we saw in Chapter Seven, some approaches take agency, self-governance, and autonomy to be constitutively tied to the pursuit of what is good or morally required. On Kantian views, for instance, action requires autonomy, which in turn requires that we choose our actions in conformity with and out of respect for the moral law. Such approaches seem to have the

counterintuitive implication that our morally bad actions are less fully autonomous than our good ones. This in turn poses complications for holding people fully responsible for their bad actions, which surely we sometimes wish to do.

In a series of important essays, Harry Frankfurt develops an alternative understanding of what it is for a person's will to be fully her own. A central commitment of Frankfurt's approach, in contrast with that of the Kantian, is that moral goodness and autonomy are independent of one another. There is no intrinsic obstacle on his view to evil characters like Iago or Caligula being fully self-governed and capable of action *par excellence*. The extent of their autonomy depends not on whether they act rightly, but on whether they act in a way that aligns with who they really are.

Frankfurt's approach focuses on a person's motivations or desires. He begins by pointing out that we can feel deeply alienated from our own desires, even as those desires move us to act in service of them. His classic example concerns a man who has an overwhelming desire to take a certain drug, but who experiences this desire as an external force acting on him. He desperately wishes he did not desire the drug and wants not to act so as to satisfy it. Nevertheless, the desire is strong enough that it succeeds in moving him to contact his dealer, procure the drug, and take it. By most standards, these actions are perfectly intentional. They are caused by a desire to get high and a belief that this is a way to do it. He knows quite well what he is doing and why, without the use of observation, and his bodily movements are under no one's control but his own. And yet, there is something defective about what he is doing; there is a sense in which his actions are not what he wants them to be.

Frankfurt first attempts to account for this intuition by distinguishing between "first-order" desires and "second-order" desires. The difference concerns the object of the desire – what it is that a person wants. First-order desires are ordinary desires that could have just about anything attainable as their objects – chocolate, a summer internship, world peace. The contents of second-order desires, in contrast, make essential reference to first-order desires. You have a second-order desire when you want to want something, or want not to want it. Perhaps you feel this way about doing the assigned reading for class. It would be so much easier and more enjoyable if you actually desired to do it! Thus, you desire to have the desire to do the reading, even though you may or may not actually desire to do it. Intuitively, the idea is that we can step back, reflect on how we are actually motivated, and form desires concerning how we would *like* to be motivated. Frankfurt uses the term 'second-order volition' to pick out those second-order desires that are specifically concerned with how we would like ourselves to be moved to act, rather than just with the having of a particular desire.

The key claim is that when a person has a second-order volition in favor of the effectiveness of a first-order desire – when the student not only wants to do her reading, but wants to want to do her reading – then she is *identified* with that first-order desire. She views it as *her* desire, and something that can legitimately be her will. In contrast, the unwilling addict in the first example is alienated from what is moving him to act, as though an external force were acting on him. He has a first-order desire to take the drug, but no second-order volition in favor of that desire; indeed, his second-order desires univocally favor his (weaker) desire not to take it. Unfortunately, simply being identified with a first-order desire in this way does not guarantee that you will act

accordingly. Even with the additional motivation to abstain that is contributed by the addict's second-order desires, his first-order desire to use is just too strong. But when the agent does act from a desire that she identifies with, then that action is fully attributable to her and speaks for her. She counts as governing herself in virtue of the fact that the desires motivating her action have the right kind of hierarchical structure. Call this a "hierarchical" model of self-governance.

2. Watson's objection and Platonic alternative

Against Frankfurt's initial view of identification, Gary Watson points out that we can be equally as estranged from our second-order desires as we are from our first-order desires. After all, they're just more desires. Frankfurt had tried to allow for this possibility by conceding that identification might lie with even higher-order desires; it might be in some cases that we must form a third-order desire which takes a stand on which combination of second- and first-order desires the agent wants to be motivated by. But Watson's objection applies equally to all levels of desire: how does simply adding more desires to an agent's psychology ensure that those new desires have any special authority? As he puts the point, "… to add [second-order volitions] to the context of conflict is just to increase the number of contenders; it is not to give a special place to any of those in contention" (1975, 28). More is needed to explain why some desires speak for the agent while others do not.

Watson proposes instead that the agent's standpoint is constituted by her values. What are values? In his earlier work, what Watson had in mind are value judgments: beliefs or assessments of what is good, worthwhile, fulfilling, or defensible in life. For example, we can imagine that the unwilling addict judges that drug addiction is bad, and that it would be better to lead a clean and productive life. Watson claims that while we can be alienated from mere desires, we cannot be similarly estranged from our own value judgments. For this would require some further evaluative judgment, such as "The idea that taking drugs is bad is just a bourgeois fear of having a free and open mind; drugs are not really bad." And insofar as the addict comes to think that this latter perspective on the value of taking drugs really is the right one, then it supplants the earlier, conflicting belief. This is not to insist that there is no possibility of conflict in our value judgments, but only that there cannot be a complete renunciation of one's values without some alternative waiting in the wings. On this model, the agent is self-governing to the extent that she acts in accordance with her value judgments, and she is alienated from an attitude and the resulting actions if she takes them to be unworthy or bad. Watson calls this a "Platonic" model of self-governance, since it identifies the agent with her reasoned conception of the good rather than with the structure of her motivations.

A challenge for the Platonic model is that the scope of our value judgments seems to be both too broad and too narrow to constitute a personal perspective on how to live. For one, many of our value judgments concern things we do not personally care about and that do not motivate us. I believe that Norwegian black metal music is valuable, but I do not enjoy it and feel no impetus to listen to it. If I did find myself going to a Norwegian black metal concert, for no other reason than to listen to the music, this does seem to be an action I am in some sense estranged from.

Second, our self-governed choices and actions often seem to outstrip our beliefs about what is best to pursue in life. Beliefs are attitudes that aim to track the truth, and as such, they are subject to pressure to converge with the beliefs of other reasonable, well-informed people. If you believe that water is wet, you are committed to thinking that anyone who believes otherwise is mistaken. But when it comes to the ways of life that are central to our identities, we often do not take a single answer to be best or expect other reasonable people to converge on our choices. One might believe that the activity of doing philosophy is just as good as practicing medicine, or one might think that there is just no good way to compare the two activities and decide which is best. And in choosing to pursue philosophy as a career or a hobby, one is not thereby committed to thinking that other reasonable and well-informed people who do not make the same choice are mistaken. The general point is that in cases where we believe that there is no clear answer to the question "Which option is best?" – and we might think such cases are fairly frequent – we can still make self-governed choices. Our value judgments alone appear to be inadequate to account for this fact.

Perhaps, then, we should construe 'valuing' as something other than, or more than, judging to be good or valuable. Watson expresses openness to this idea in later work, though he does not specify a particular account. Possibilities include adding motivational and emotional components to valuing: perhaps in addition to judging that V is valuable, one must be motivated to act in ways that engage with or promote V, and one must care about V or otherwise be emotionally vulnerable to V. Only then can V be construed as something that one values. We will return to this idea in the section on Michael Bratman and self-governing policies. There are limits to this expansion of the idea of valuing, however. As Watson himself later acknowledges, we might wish to allow for "perverse" cases in which the agent fully embraces a course of action without seeing it as truly good or defensible. This is something the Platonic model of identification cannot do.

3. Frankfurt redux: wholeheartedness

In later work, Frankfurt adds to his hierarchical account in an effort to address Watson's objection without embracing the Platonic alternative. He grants that *merely* occupying a higher level of the hierarchy does not endow a desire with any greater legitimacy or claim on the self. Rather, this legitimacy comes with "wholeheartedness," which concerns the absence of conflict. At some point in the hierarchy, we will reach a level at which there is no further tension between the agent's desires. Let's imagine that Vanessa, like most of us, desires both to be healthy and to enjoy unhealthy pleasures. Further, she has a second-order desire in favor of being motivated to act in healthy ways and no competing second-order desire that favors being motivated by unhealthy pleasures. The fact that Vanessa identifies with her desire to be healthy consists in the presence of the second-order-desire, but the fact that she identifies with her second-order desire consists in her unambivalence: at that level, there is no disunity in what she wants.

In his initial characterization of wholeheartedness, Frankfurt required that the agent arrive at this conflict-free state via the making of a decisive commitment that "resounds" through all the higher levels. But in later development of the notion, he drops the idea of a resounding commitment and speaks instead of a kind of reflexive satisfaction with one's psychic state.

Satisfaction with an unambivalent hierarchy of attitudes involves the absence of a desire to change one's condition, but it need not entail the belief that one's condition is satisfactory or that the desires one identifies with are better than those one rejects. Vanessa might believe health is more valuable than unhealthy pleasure, but it is just as possible to identify oneself with self-destruction while believing that it would be better to be fit and well. What matters fundamentally is wholeheartedness itself, not the reasons behind it.

In the most central cases, on Frankfurt's later view, it is not only that the agent is wholehearted about a desire. She also finds herself unable not to be, or even finds it unthinkable not to be. Frankfurt does not have in mind a mere compulsion, but what he calls a "volitional necessity," often stemming from love. For example, a parent might find it volitionally impossible to disown his criminal adult child even though he recognizes that there are good reasons to do so. These volitional necessities place limits on what we can do, but they also give the self a determinate shape, thus allowing for the very possibility of self-governed action. Wholeheartedness is important not just because it helps to make sense of identification, but because ambivalence is a threat to the very existence and preservation of the self.

4. Bratman on self-governing policies

Bratman articulates a third conception of self-governance that draws on the resources of his planning theory of intention, as discussed in Chapter Five. Bratman's approach is a non-Platonic version of the appeal to valuing, meaning that it is not committed to the idea that there is an objective Good that should determine what we value in life. Instead, he proposes that there is a form of valuing that consists in a kind of intention rather than a belief or a desire. More specifically, the focus of his account is on what he calls "self-governing policies," where a policy is a general intention that applies across a variety of particular cases. A self-governing policy is a general intention to give weight to certain considerations in practical reasoning and action – to treat those considerations as reasons, and to accord them some determinate amount of significance. For example, you might have a self-governing policy to treat considerations bearing on academic success as reason-giving in deliberation and action, and in particular, to treat them as more important than considerations bearing on your social life. You might also treat your social life as important, and take yourself to have reason to meet up with friends when there is no competing academic demand. But to the extent that your policies give more weight to school than to socializing, this means that you value school more. And to the extent that your policies give no weight at all to the violent rages that sometimes strike when you are driving in traffic and haven't yet had your coffee, this means that the propensity to traffic-related violence is external to your agential standpoint altogether.

Our self-governing policies will often be informed by our judgments about what is independently valuable, but they are not fully constrained by them. Our policies may even conflict with our value judgments, though this would likely amount to a rationally criticizable form of incoherence in one's overall perspective. Most importantly, we can form self-governing policies even when we do not believe that there is a uniquely correct answer about what is best, and without

expecting others to converge on the same commitments. It is possible to weigh academic success more heavily than socializing in your own deliberation while thinking that your friend who does the opposite is making an equally reasonable choice.

Like the other two views, Bratman's account owes an explanation of why self-governing policies have a kind of authority, such that they are the right kind of psychological state to constitute the agent's true stance. His answer appeals to the role that intentions and policies play in coordinating thought and action, not only at a particular time, but also over time. It is central to our lives that we frequently plan ahead for the future, and that we carry out intentions that were formed in the past. Policies in particular play an important role in organizing our lives over time in virtue of aiming to settle questions about how to reason and act not just on one occasion, but on all similar occasions. And by playing this cross-temporal role, they provide a kind of continuity in the agent's psychology in virtue of which she counts as being the same person over time. Compare this claim with an influential theory of personal identity over time in which memory plays the role of tying disparate episodes of consciousness together into a sufficiently continuous stream (this kind of view is usually called 'Lockean', though it is not quite John Locke's own view). Roughly, if you can remember having some conscious experience, then you are the person who had that experience. Similarly, Bratman's proposal is that intentions and policies help to give us a unified diachronic identity: roughly, among other things, your future self is the person who will carry out the plans you make now, and your past self is the person who made the plans that you are now guided by. And this role is what gives them the authority to "speak for the self," such that when they guide action, you are governing yourself.

5. Skepticism about self-governance: a genealogical worry

One of the major sources of appeal for these theories is that they seem to allow for the possibility of reflectively choosing what kind of person to be. Though they differ in the details, all three views emphasize the importance of stepping back from the appetites, lusts, rages, jealousies, and other psychic elements that motivate us and asking "Is this really me?" The idea that we can to some extent determine who we are is connected to the vexing issue of moral responsibility. According to our commonsense way of thinking about it, a person can only be responsible for what she does – blameworthy for her bad actions and praiseworthy for her good ones – if those actions are attributable to *her*. If she does something bad because she was coerced into doing it, or compelled by some external force, then this is generally considered an excuse. But the worry then arises: is anything really attributable to me? If I'm just another material object, governed by the laws of physics, born with a certain nature, and so forth, how can I be responsible for anything that I do? This worry seems to be assuaged somewhat by the thought that we can reflectively assess the motivations we start with, disavow some, and endorse others, thereby actively shaping the person that we are.

However, we might see this solution as merely kicking the can down the road. After all, who is it that is doing the reflecting? The process of reflectively forming higher-order desires, policies, and values must be guided by something, on pain of being completely arbitrary. But

especially at the outset, the point of view from which we reflect will itself be shaped by external forces, such as the values and policies that our parents, teachers, and community had as we were growing up. Susan Wolf offers us the example of JoJo, the son of an evil and sadistic dictator. JoJo was educated by his father and allowed to accompany him as he went about his daily business of oppressing and torturing. He idolized his father, and as a result developed values and higher-order desires that are very much like the ones his father had. As an adult, he behaves in the same sadistic ways that his father did, and he is reflectively wholehearted about it. As Wolf describes the case, "When he steps back and asks "Do I really want to be this sort of person?" his answer is a resounding "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal" (1987, 379).

Is JoJo a self-governing agent whose actions are "full-blooded?" Once we see the explanation for how he became the person that he is today, we are tempted to say no. We can take the challenge a step further by imagining that a hypnotist or neuroscientist simply implants a higher-order desire, self-governing policy, or value judgment into an agent, after which it goes on to play its normal functional role in coordinating reasoning and action. Would the effects of this attitude still count as the agent governing herself? According to the theories in question, it is unclear why not, since those theories do not impose any strict requirements on the genealogy of the attitudes that constitute us as autonomous agents.

We could, of course, add some kind of genealogical constraint on where an autonomous agent's standpoint must come from. Alternatively, proponents of these views might counter that the genealogy of JoJo's commitments is not the fundamental problem in his case. Though Wolf herself was interested in moral responsibility rather than self-governance, we might for instance adopt her proposal that the attitudes constituting the agent's standpoint must be *sane*. While none of us can avoid being shaped by forces outside of us, a sane person – unlike JoJo – has the capacity to reflectively alter her guiding commitments in light of new experiences and new information. Attitudes, policies, and values that were instilled by external forces will not persist in such a person in ways that are insensitive to his other commitments, or to what he may learn when exposed to a new environment. Perhaps we cannot create ourselves out of nothing, but a sane person can revise himself, and that may be enough.

6. Self-governance and plan rationality

In addition to a connection with being held responsible for our actions, we might also think there is a connection between self-governance and rationality. Recall from Chapter Five that intentions and plans are subject to certain rational requirements, such that if a person violates these requirements, there is a respect in which she is being irrational. One such requirement concerns intending the means believed necessary to your intended ends. If you are "means-end incoherent," in that you intend to achieve some goal E, believe that you will not achieve it unless you now intend means M, but do not intend M, then you are in an irrational state. Similarly, if you have multiple intentions that you believe are not mutually compatible, in that carrying out one would rule out realizing the other, then you are irrationally inconsistent. Third, more controversially, there might

be something irrational about being "unstable" in one's intentions over time – abandoning old intentions and forming new ones without good reason for changing one's mind.

But now the question arises of why we should even care about being irrational. When we call someone 'rational' or 'irrational', we are describing patterns in their mental lives. Why, though, should that person prefer to exhibit rational patterns rather than irrational ones? Is there any good reason to care about satisfying these "requirements" of rationality? After all, there are plenty of norms that a person can reasonably choose not to care about, and thus not even try to satisfy. Good etiquette requires that one never place one's elbows on the dinner table, and being fashionable requires that one never wear socks with sandals. Violating these norms makes it less likely that you will be invited to dinner parties. But if a person is informed of these facts and simply does not care about being polite or fashionable, there is no obvious mistake he is making in behaving and dressing tastelessly. In contrast, when we charge a person with irrationality, we do not think it optional that he should agree that this is a mistake and try to adjust to a more rational state.

Bratman argues that the rational requirements on planning express conditions on being self-governed. Let us think of being self-governed in an atomistic way, as relative to a particular practical question like "Whether or not to end the friendship?" This allows us to say, plausibly, that agents are often self-governed with respect to some parts of their practical lives but not others. In order to be self-governed with respect to a practical problem, on Bratman's approach, there must be a coherent answer to the question "Where does the agent stand?" on that issue. Does Liz stand on the side of cutting off a toxic relationship, or does she stand on the side of continuing it? Suppose now that Liz has inconsistent intentions on the subject; she intends to end the friendship immediately, but she also intends to continue seeking her friend's company and confiding in her. There seems to be no fact of the matter as to where she stands, since in taking both sides, she succeeds in taking neither. Similarly, suppose she intends to end the friendship and believes that she will not accomplish this unless she cuts off all contact. If she fails to form the latter intention, she fails to have a coherent standpoint – she is both committed and not committed to the breakup. And third, if her intention to leave the relationship is unstable over time – she intends to leave today, and then abandons that intention tomorrow, only to resume it the next day – then we might see her as lacking a coherent standpoint *over* time, even if she has one at each point in time.

If this is right, then an agent who violates any of the rational requirements on intending fails to be self-governing in that domain. Even if she does end up acting, either by ending the friendship or continuing it, the action is not fully attributable to a standpoint that speaks for her. Thus, if she asks what reason she has to be rational rather than irrational, one answer is "Because you will fail to govern your own life in that respect if you have rationally impermissible intentions!" Of course, this raises the further question of whether and why we must care about being self-governed, or whether it is possible not to be concerned with that objective. This worry recalls the "shmagency" objection to constitutive-aim theories that we encountered in Chapter Seven. But if we can assume that most people *do* care at least to some extent about governing their own lives, then this concern supports a concern for being rational in our intentions. Importantly, this is not meant to be the only justification for adhering to the requirements of intention rationality. We can

also point out that following them will generally make us more effective in realizing our goals over the long run by facilitating coordination and preventing us from stepping on our own feet. Considerations of self-governance are meant to complement this kind of pragmatic justification, not to supplant it.

Summary

The driving thought behind this chapter is that even when an action is intentional, the agent may or may not identify with it as the true expression of her will. We sometimes perform intentional actions "in spite of ourselves," while other actions of ours speak to who we really are. Views like Frankfurt's and Bratman's both attempt to account for this intuition by locating identification or self-governance in the hierarchical structure of the person's will. Frankfurt locates this structure in the person's desires, while Bratman locates it in self-governing policies, but the idea in both cases is that certain aspects of the person's psychology have a kind of "agential authority:" when they function to guide reasoning and action, the agent is governing herself. This kind of view makes agential authority independent of content, since in principle, the relevant desires or policies could be about anything at all. In contrast, views like Watson's tie agential authority and self-governance to the agent's values. This kind of account is not hierarchical, since the authority of the agent's values to speak for her does not derive from their place in a given psychological structure. Rather, it derives from what they are about: they speak for her because they represent her view of what is good or worth doing.

Skepticism about this project may arise for broadly the same reason that people are skeptical about the existence of free will. About any given psychological structure or set of attitudes that are said to speak for the agent, we can ask "Where did that structure/those attitudes come from?" Even if we privilege attitudes that arise from, say, reflective deliberation, commitment, or love, it will usually be possible to show that these attitudes were heavily shaped by factors outside the agent – her upbringing, other influences of her environment, and so forth. Thus, we might worry that none of our actions ever truly amounts to being *self*-governed.

Suggested Reading

The problem of identification and self-governance is developed by Frankfurt over a series of essays that can be found in the collections *The Importance of What We Care About* and *Necessity, Volition, and Love*. Especially important papers in the former volume include "Freedom of the Will and the Concept of a Person," "Identification and Externality," and "Identification and Wholeheartedness." In the latter volume, "The Faintest Passion" and "Autonomy, Necessity, and Love" are especially relevant.

Watson's criticism of Frankfurt's initial view is posed in his essay "Free Agency." He revisits and modifies his Platonic position in "Free Action and Free Will." Both can be found the *Agency and Answerability* collection.

Bratman's views on this question are developed in essays that can be found in *Structures of Agency* and *Planning, Time, and Self-Governance: Essays in Practical Rationality*. His paper "Three Theories of Self-Governance" is especially helpful in directly contrasting the views presented in this chapter.

The example of JoJo the dictator comes from Wolf's paper "Sanity and the Metaphysics of Responsibility." Though she is explicitly concerned with moral responsibility rather than self-governance in that paper, many of the ideas there and in related work apply also to the topic of attributability.

9. Temptation, Weakness, and Strength of Will

Anna Karenina, let us suppose, sincerely judged that she should not have an affair with Count Vronsky. She knew that if she did, her politically powerful husband Karenin and the moral norms of Russian society would ensure that she would become a social outcast and lose access to her son Seryozha. It would also injure her young friend Kitty, who was herself in love with Vronsky. Though Anna was attracted to him, she judged that this was insufficient reason to proceed with the affair given all the considerations weighing against it. And yet, she did it anyway.

We sometimes do questionable things because we are reckless. We throw caution to the wind and choose, however imprudently, to have affairs and accept the consequences. In other cases, we are the victims of compulsion. Compulsion renders a person unable to make a real choice. In such cases, even if there is a voluntary act involved, it is primarily attributable to some force acting on the person and not to him. Some compulsive forces are literally external to the person, as when a mugger demands one's wallet at gunpoint. Other such forces are internal. Compulsive psychiatric disorders can result in behaviors like excessive hand-washing that have many of the hallmarks of voluntary action, but that are not under the agent's control in the relevant sense. The obsessive-compulsive agent must wash her hands again whether she wants to or not, indicating that the behavior is attributable to the disease rather than to her. And motivational afflictions resulting from depression or *accidie* (a kind of spiritual listlessness or torpor) can block the normal effect of the agent's judgments about what it is best to do, leaving her unable to leave the couch even though she knows it would do her good to go for a walk outside.

Anna's case is much more puzzling than these. There is a sense in which she loses control over herself when she goes forward with the affair, but there is another sense in which she has full control over what she was doing. Although her options were constrained by her social situation, she genuinely had a choice between being loyal to her family and proceeding with the illicit romance. At the same time, she was not simply reckless; she is acting against her sincere judgment that she should not have the affair. Her choice to have an affair exhibited what is often called "weakness of the will," "incontinence," or following the ancient Greeks, "akrasia." Unlike cases of coercion or psychiatric compulsion, weak-willed or akratic actions are performed freely, in the sense that they are attributable to the agent and not to some force acting on her. But like most cases of coercion and compulsion, akratic actions are not performed because the agent judges that they are the best thing to do. Rather, she defies her own best judgment, intentionally doing one thing while taking some other available option to be better. What is going on in such cases?

Before examining various answers to this question, let me introduce two clarifications to the puzzle. First, there are in fact two distinct versions of the problem of weakness of will. In one version – the "synchronic" version – we observe the agent at a single point in time at which she simultaneously judges action A to be better than action B (where A might simply be refraining from doing anything) and intentionally performs action B. We can imagine Anna thinking to herself "This is wrong, the best thing to do would be to stay home!" even as she steps out the door to meet

Vronsky. In the second version – the "diachronic" version – the agent never chooses to act in a way that conflicts with her own best judgment at that time. However, her judgment about what is best fluctuates over time, leading her to choose in a way that conflicts with a previous and perhaps future judgment about what she ought to do. The diachronically weak-willed Anna judges in the morning when she is playing with her son that she should break off the affair, but changes her mind when evening comes and judges that continuing to see Vronsky is the better option. The next morning, her judgment again reverses itself and she regrets her decision. It will be important to keep the synchronic and diachronic versions of the puzzle distinct, since it may be that they admit of different explanations.

Second, it is tempting to suppose, as many philosophers throughout history have done, that weakness of will must always be a matter of some base appetite or emotion getting the better of the agent's reasoned sense of duty or morality. This may be right, but there are powerful reasons to think that weakness of will need not always involve doing the thing considered to be less moral or prudent. Consider the following example offered by Donald Davidson, in which the agent acts weakly out of a concern for duty against the greater acknowledged value of sensual indulgence:

> I have just relaxed in my bed after a hard day when it occurs to me that I have not brushed my teeth. Concern for my health bids me rise and brush; sensual indulgence suggests I forget my teeth for once. I weigh the alternatives in the light of the reasons: on the one hand, my teeth are strong, and at my age decay is slow. It won't matter much if I don't brush them. On the other hand, if I get up, it will spoil my calm and may result in a bad night's sleep. Everything considered I judge I would do better to stay in bed. Yet my feeling that I ought to brush my teeth is too strong for me: wearily I leave my bed and brush my teeth. My act is clearly intentional, although against my better judgment, and so is incontinent (1969a, 30).

This kind of example suggests that weakness of will is not necessarily a matter of failing to do what you know to be the moral or prudent thing; rather, it is a matter of acting against your own best judgment, whatever it happens to recommend. Davidson concludes that "incontinence is not essentially a problem in moral philosophy, but a problem in the philosophy of action" (1969a, 30).

1. Is synchronic akrasia even possible?

Akrasia is such a puzzling phenomenon that Plato's Socrates held it to be impossible. In the *Protagoras*, Socrates makes the following claim:

> Now, no one goes willingly toward the bad or what he believes to be bad; neither is it in human nature, so it seems, to want to go toward what one believes to be bad instead of to the good. And when forced to choose between one of two bad things, no one will choose the greater if he is able to choose the lesser (358c-d).

Socrates concludes that whenever a person fails to pursue what is in fact the best available option, it is due to ignorance of what is really best. The thought is that in the moment, at least, Anna must believe that seeing Vronsky is in fact her best option. If she had believed that loyalty to her family was best, she would have wanted to end the affair more than she wanted to continue it. And if she was free to choose, surely she chose what she most wanted to do. This does not rule out diachronic akrasia, but according to the Socratic view as it is commonly understood, we necessarily do what we believe at the time to be best if we do anything intentionally at all.

As Aristotle remarked, the claim that no one errs willingly conflicts with the common wisdom. In this day and age, how could it be that all those who smoke cigarettes are ignorant of how bad it is to smoke, or that no one really wants most to quit smoking even at the moment he lights a cigarette? Although Aristotle sketched a view in the *Nicomachean Ethics* that is similar to the Socratic thesis, he attempted to accommodate the appearance of genuinely akratic actions. He argued that the incontinent person does know, in a sense, that he should be doing something other than what he does, but in another sense does not know. Under the influence of passion, he is like a person who is drunk, mad, or asleep – he has knowledge of what is good, and can utter words like "I really shouldn't be doing this," but his knowledge is ineffective and therefore possessed only in a qualified way.

Aquinas largely followed Aristotle in his explanation of incontinence, helpfully elaborating on the way in which the incontinent person might have abstract, universal knowledge of what is good in principle while failing to make the transition to particular knowledge about what is good to do here and now. In other words, the smoker knows that "things that are unhealthy are bad and to be avoided," but fails to know in the moment that "smoking this particular cigarette is bad." Unlike Aristotle, however, Aquinas gives a role to the faculty of the will (*voluntas*). Aristotle held that when the incontinent person acts, he acts against his own choice (*prohairesis)*. This creates a mystery about why the resulting act should be considered voluntary and something he can be held responsible for rather than an instance of compulsion. On Aquinas's view, a person can only be held responsible for acts that he willed, and so it must be the case that incontinent agents will their actions. And since he held the view that willing must take place under the Guise of the Good (see Chapter Seven), it follows that the incontinent person must judge in the moment that his particular act is good even if he knows that in general such acts are not good.

Although there is some variation in the details, each of these views denies the possibility of fully clear-eyed synchronic akrasia. The underlying reason is that they all subscribe to the Guise of the Good thesis, and thus take motivation to be directly and necessarily connected to evaluative judgment. Socrates and Aristotle put the connection in terms of the motivational power of desire, whereas Aquinas takes the connection to hold between evaluative judgment and the motivation of the will. But in each case, the thought is that intentional action is essentially a form of rational activity, motivated and guided by a person's conception of what is good. If Anna truly judged that she ought to refrain from having the affair, without some defect or interference clouding her judgment, the only explanation of her failure to refrain is that she was not free to do so – she must have been involuntarily compelled to do otherwise. So either her "judgment" did not embody her

true assessment of what is best, or her "choice" to have the affair was actually the result of compulsion.

Those who deny the possibility of akrasia are resourceful in arguing that apparent counterexamples are actually cases of compulsion or insincere judgment. We will see in the next two sections why we would be pushed in the direction of seeing such actions as compulsive. With respect to arguing that the relevant judgment must be insincere, there are several possibilities. Anna might be exhibiting a kind of "backsliding," in which she changes her judgment about what she ought to do in the moment before she acts. Alternatively, she might be using the concept 'ought' in what R.M. Hare calls an "off-color" way, to refer to what is conventionally expected or to express that she would refrain from the affair if she could, though she can't.

2. A failure of reasoning?

And yet – the conviction that there really are actions that are intentional, free, and akratic is very difficult to shake. Davidson usefully frames the puzzle in terms of three principles, each of which has a great deal of initial plausibility but which appear to contradict one another:

> P_1. If an agent wants to do X more than he wants to do Y and he believes himself to be free to do either X or Y, then he will intentionally do X if he does either X or Y intentionally.
>
> P_2. If an agent judges that it would be better to do X than to do Y, then he wants to do X more than he wants to do Y.
>
> P_3: There are incontinent actions.

Davidson's own solution is to distinguish between two different evaluative judgments the agent reaches in cases of incontinent action. On his view, to will or intend an action is to reach an all-out, unconditional judgment in favor of performing that action (see Chapter 5, section 3c). In other words, the judgment is not relative to some set of considerations, as when one judges "*If* it's raining, other things equal, I ought to take my umbrella with me." It does not follow from this conditional judgment that I ought to take my umbrella, since it might not be raining, or there might be other reasons not to take the umbrella that outweigh the significance of the rain (perhaps it is a valuable heirloom). An intention to take one's umbrella, in contrast, is a judgment with the simple form "I ought to take my umbrella" – no conditions, qualifiers, or escape-clauses. With this distinction in hand, we can re-interpret P_2 as referring specifically to the latter, unconditional kind of judgment: "If an agent judges *unconditionally* that it would be better to do X than Y, then he wants to do X more than Y."

In cases of akrasia, the agent forms two judgments. In addition to the unconditional judgment embodying her intention, she also reaches what Davidson calls a conditional but "all-things-considered" judgment: "Relative to all of the relevant considerations, I ought not to take

my umbrella with me." This judgment constitutes the agent's "best" judgment, since it takes into account all the reasons bearing on what she ought to do, including whatever reasons informed her all-out unconditional judgment. However, it is still conditional and thus not the expression of the agent's will; it does not determine what she does. P_3 – the possibility of incontinent actions – is thus vindicated, since the agent can freely and intentionally act against her all-things-considered judgment. Anna judges that it is all-things-considered best to refrain from having the affair, but she intentionally has the affair because she unconditionally judges that it is best to have the affair. It is the latter judgment that is connected with "wanting most," ensuring the result that Anna wants most to have the affair and therefore intentionally does so.

Davidson's view may succeed in showing how we can hold onto the thought that there is an essential connection between motivation and evaluative judgment without making akrasia into an outright logical contradiction. However, it seems to leave much of the mystery firmly in place: why on earth would anyone reason to the conclusion that she should perform action Y after having determined that X is all-things-considered better than Y? We would not expect such an obvious mistake in reasoning to be made so frequently and repeatedly. Davidson gestures at the possibility of various psychological explanations of the phenomenon, but denies that the agent herself can have any reason for doing Y rather than X: that choice is "essentially surd." This may strike us as unsatisfying with respect to the task of distinguishing akrasia from compulsion, for if the agent cannot make any rational sense of her choice to do Y rather than X, it is hard to see how she can view herself as having been free to X rather than as the victim of an overwhelming urge. Further, one might complain that Davidson's view vindicates the letter but not the spirit of the claim that our motivation to act is tightly connected with our judgments about the good. Why should it be that motivation necessarily accompanies our unconditional evaluative judgments rather than our all-things-considered judgments, given that the latter has a better claim to represent our conception of the good?

3. A divergence between evaluation and motivation?

Dissatisfaction with the conclusion that akrasia is either impossible or the result of obviously defective reasoning may lead us to question the idea stated in P_2: that motivation is tightly connected to evaluative judgment in agents like us. In a sense, Davidson's solution tries to loosen this connection while at the same time retaining P_2 by adopting a counterintuitive understanding of what we mean by 'evaluative judgment'. A more intuitive version of this strategy simply rejects P_2. We can grant that a *fully rational* agent is necessarily motivated by her judgment about what it is best to do, all things considered, but point out that akrasia is a kind of irrationality. If it is possible to desire strongly to do something without taking it to be an especially good thing to do, or for one's evaluative judgments to have little or no motivational impact, this might explain how akratic action occurs.

For example, conditions like depression and accidie can render a person unmotivated to pursue what she knows to be good, even when it comes to her own well-being. J. David Velleman asks us to imagine a person suffering from depression who decides to smash some crockery,

precisely because it is a completely worthless act (not to feel better – as Velleman points out, "someone who smashes crockery in order to feel better didn't feel all that bad to begin with" (1992a, 20-21). This kind of case seems to show that it is possible to be motivated to perform an action that one does not judge to be good in any respect, let alone the best available option. It is even easier to imagine that such a person is capable of judging that she really should go for a walk and eat a nutritious meal without this judgment moving her in the way that it would if she were rational. There is thus good reason to be skeptical of P_2.

It may seem that this is all we need: if motivation to act can be out of proportion with our evaluative judgments, plenty of room is left for akratic actions. However, there is more philosophical work to be done in order to successfully differentiate weakness from compulsion. The weak agent could have resisted her desire but did not, whereas the victim of compulsion could not have resisted. The skeptic about akrasia might grant that P_2 is false while insisting that whenever motivation and evaluation come apart, it is because the agent is the victim of some compulsion or other pathology that undermined her agency and rendered the better option in fact unavailable to her.

The argument for this is that there seems to be no explanation of *why* the weak agent did not resist temptation if (1) she could have resisted, and (2) she truly judged that it would be better to resist. If Anna *chose* not to exercise sufficient self-control to resist her desire to see Vronsky, knowing that self-control would be necessary for implementing her judgment that she should stay home, then she seems simply to have given up that judgment (unless she akratically fails to exercise self-control, in which case a vicious regress looms). On the other hand, if she knew that self-control was necessary and did not choose to forgo the use of it, the only explanation seems to be that she lacked the capacity to resist and so is compulsive rather than weak. To accommodate this dilemma while avoiding the conclusion that there is no such thing as akratic action, Gary Watson proposes that the akratic action differs from the compulsive action only counterfactually. At the time, the weak-willed agent could not have resisted any more than the compulsive agent, but unlike the compulsive agent, if the weak-willed agent had had a normal amount of self-control (defined by the capacity of others in her community), she would have been able to resist. She counts as weak-willed because she has failed over time to develop what is communally considered to be a sufficient amount of self-control.

The difficulty of making good on the apparently simple strategy of rejecting P_2 highlights what a vexing puzzle akrasia poses. The Socratic view that there is no such thing as clear-eyed akrasia can strike us at first as a rather mad denial of the obvious. But examination of the alternative helps to make clear why one would be drawn toward that view. To explain an occurrence as an exercise of agency is to show how the agent had control over what happened. And an attractive way of understanding what it means for an agent to have control appeals to her conception of what is good: an occurrence was an action and thus under her control if it is explained by the fact that she judged it to be the best (or a sufficiently good) thing to do. If it happened in spite of her judging that it was not a sufficiently good thing to do, the challenge is to explain why it should count as an exercise of her agency at all.

4. Is akrasia necessarily irrational?

Let us now take a step back. Those who reject the Socratic view are committed to thinking that it is possible to act freely and intentionally without the sanction of one's judgment about what is best. Once this space has been opened up, it raises a further question: is it necessarily *irrational* to do so? Or can it be rational to act against your own best judgment? Now, all can agree that acting against your own best judgment will sometimes produce what is in fact the best result, simply because our judgments about what is best can be in error. Think of George Costanza, from the TV show *Seinfeld*, having the epiphany that every instinct he has in every aspect of life is mistaken. He would have been better off if he had each time done the opposite of what his poor judgment had led him to do. Insofar as George ever acted akratically prior to his epiphany, he would have ended up performing what was in fact the better action. But the interesting question is whether he could have been rational in doing so.

Much depends, of course, on what we mean by 'rational'. On one possible "externalist" understanding, an action is rational just in case it is the right thing to do, no matter what the agent desires or believes. For example, let us assume that torture is impermissible no matter what anyone believes about it. According to the externalist conception of rational action, a military officer who wholeheartedly believes that torturing prisoners of war is morally right would be acting rationally if he weakly succumbed to the temptation not to torture his prisoner. This is just another way of saying that in declining to torture his prisoner, he got the right result. But again, the more interesting question is whether he is being rational from a perspective that is internal to what he believes and desires.

It is plausible to insist that such an agent could not be fully rational, in that his action and his evaluative judgment are mutually inconsistent. In so acting, he must be something of a mystery to himself. That said, as Nomy Arpaly argues, it does not necessarily follow that he would be *more* internally coherent if he managed to act in accordance with his best judgment. It might be that the officer does generally value human dignity and desire not to inflict pain, and that his judgment in favor of torturing prisoners is in fact the outlier in his psychology. His reluctance to proceed with the torture might actually be a manifestation of his tacit recognition of the prisoner's humanity. If so, refusing to torture might be what he has most reason to do even from the point of view of his own desires and values, while his judgment to the contrary is simply mistaken. To insist that he would be more rational if he were nevertheless to act in accordance with that judgment would be to ignore the relevance of the officer's other attitudes; after all, as Arpaly puts it, "an agent's best judgment is just another belief" (2000, 512).

Much of the debate over whether akrasia could be rational revolves around the question of what temptation is. On Aristotle's view, incontinent actions are construed as the result of intemperate appetites for the pleasures of taste and touch. Since intemperance is always a vice, an intemperate appetite could never lead the agent to act rationally (or as Aristotelian views would have it, virtuously) against her own best judgment. Similarly, according to the "Scholastic" view associated with Aquinas, to suffer from temptation is akin to undergoing a perceptual illusion, like seeing a mirage. To be tempted is to have a desire to do something, but on this kind of view, desire

is a cognitive state that represents the world as being a certain way, much like perception: to desire a cold glass of lemonade is (at least in part) to represent the drinking of the lemonade as good. But like the relation that perception stands in to reflective belief, desires are generally spontaneous, stimulus-driven, and potentially illusory, such that a more detached and reflective judgment may be needed to override their influence. Perception can make it appear to one as if there is standing water on the highway, but an experienced perceiver will override this appearance and judge that there is in fact no water. Likewise, an agent to whom the glass of lemonade *appears* to be good might nevertheless judge that this appearance is deceiving (because of the high sugar content, say) and that all things considered, it would not be good to drink it. If temptation is a matter of a mere appearance of the good conflicting with a reflective judgment about what is truly good, where the grounds for the reflective judgment include but are not limited to the appearance of goodness, then the reflective judgment will always have more rational authority than the opposing temptation.

Kantian views have different grounds for holding that reasoned judgment must always be more authoritative than inclination or desire. On this kind of view, there is no sense to be made of an action being the right thing to do independently of the procedure by which the action was chosen. And this procedure must involve the exercise of reflective judgment, since the agent must judge her maxim to be in conformity with the Categorical Imperative. Like the Socratic view, then, Kantian views leave no room for a behavior to count as a fully autonomous action if it is not ratified by the agent's reasoned judgment. If an inclination directly causes the behavior without the participation of reason, then the behavior is arguably not an action at all, or is at most heteronomous rather than autonomous. And if reason does somehow participate in the production of an action that conflicts with it, this is clearly a defective form of willing, and could not be more rational than acting in accordance with reason.

In contrast, on the more Humean conception of rational agency to which Arpaly is sympathetic, reflective deliberation and evaluative judgment play no essential role in transforming mere behavior into full-blooded autonomous action. Deliberation and judgment can be helpful in *figuring out* what to do, but they have no more status than the agent's desires when it comes to *making* an action the thing to do. On this kind of view, there is nothing inherently defective about being motivated directly by a desire – indeed, only desires can motivate us on a purely Humean view – and nothing inherently better about an action's being sanctioned by an evaluative judgment. In fact, given the cognitive limitations of normal human thinkers and the many considerations bearing on what to do, it is no surprise that a given episode of deliberation can get it wrong and produce a mistaken judgment about what you ought to do. If temptation is simply a desire that conflicts with an evaluative judgment, that desire may well accord better with what the agent has most reason to do than the judgment does.

5. Weakness of will over time

Return now to the distinction we drew earlier between synchronic and diachronic weakness of will. The basic idea was that in the former case, the action in question is in conflict with a judgment that is held at the time the action is performed, whereas in the latter case the action and

the conflicting judgment occur at different times. That is, the weakness lies in the fact that under the influence of temptation, the agent temporarily changes her mind about how she ought to act. As mentioned earlier, the possibility of diachronic akrasia might help to explain why Plato's Socrates was willing to deny what seems to be an obviously possible phenomenon. That is, the Socratic view allows that while synchronic akrasia is impossible, an agent can believe at time t_1 that she should not continue the affair, only to change her mind in favor of concupiscence at t_2 once her lover is in front of her. Such an agent is also disposed to reacquire her abstemious belief at time t_3 and regret that she indulged at t_2. If most cases of what look like synchronic akrasia are in fact caused by evaluative judgments that are unstable over time, this might help to explain the premium that Socrates placed on knowledge, since he viewed knowledge as more stable over time than mere belief.

Richard Holton argues that the ordinary, non-philosophical conception of weakness of will is actually the diachronic rather than the synchronic phenomenon, and that the agent's best judgment need not even play a role in explaining why the action is weak. Rather, we exhibit weakness of will whenever we too readily abandon the intentions or resolutions we have previously formed. Holton follows Michael Bratman in understanding intentions to be plan-states that are irreducible to belief, desire, or some combination of the two (see Chapter Five, section 4). It follows that if Esther forms the intention to have only one glass of wine at the party tonight, she need not have judged that having one glass is her best option (she might not have formed any judgment, or she might be indifferent between one glass and two). Nevertheless, if she abandons her intention at the party as a result of temptation and proceeds to have two glasses of wine, there is a kind of weakness she has exhibited across time. Holton emphasizes that this is especially the case if Esther resolved on one glass of wine precisely because she anticipated that she might be inclined to drink more than that once she arrives at the party.

This is an extremely plausible thought, but it is surprisingly difficult to pin down exactly where the irrationality lies in these cases. One suggestion might be that although Esther does not actually judge at the time that she ought only to have one glass, she would reach that judgment if she did deliberate about it. The problem is that this is simply not true in many cases. As the Aristotelian picture suggests, temptation often induces what Holton calls "judgment shift," in which the agent's evaluative ranking of her options is temporarily reordered. Esther not only wants to drink the second glass of wine; the temptation to do so causes her to judge in the moment that drinking more wine is the best thing to do, or would cause her to make this judgment if she were to deliberate.

A second suggestion might be that even if Esther is such as to judge at the party that more wine is better than less wine, she would be irrational in so judging, and the irrationality of drinking the wine is grounded in this fact. But this claim is problematic as well. If she desires most to drink the wine, even anticipating that she will suffer for it the following day, why would it be irrational for her to judge that she ought to go ahead and do it? Her desires about her own experiences seem to be most relevant to what she has reason to do in this case. We might try saying that we should always follow through with our resolutions, even when they conflict with what we now believe we ought to do. This is madness, however; some resolutions we form are ill-considered even at the

time, and over time, circumstances can change and new information can come to light. It is clear that we should sometimes abandon our plans rather than stubbornly following through. Nor can we simply say that we should never abandon a resolution in situations where the present circumstances – the conviviality of the party, say – induces a change in our desires. Sometimes being vividly confronted with the relevant circumstances, becoming emotionally engaged, and so forth is precisely what is needed to find out what you truly want.

Finally, we might appeal to the thought that the agent is disposed to regret her choice in these cases to explain why the action is rationally problematic. Esther's decision to drink more than one glass of wine at the party conflicts not only with her previous resolution, but also with the feeling of regret she will likely have on the following day. This has the potential to distinguish between weakness of will and genuine changes in the agent's view about what is best to do, since she is not disposed to regret the latter kind of change. Still, not all weak-willed agents are disposed to regret their choices. For instance, some people actively cultivate a "no regrets" policy toward their past decisions, even ones that were the product of temptation. It is therefore unclear that the appeal to future regret will suffice to explain what is defective about the agent's choice in these cases, though it might give us some defeasible evidence.

6. Self-control

The counterpart to weakness of will is self-control. What is it to have the capacity for self-control, and what is it to exercise self-control in order to overcome temptation?

Many of the techniques we use to control ourselves are similar in kind to those we can use to control other people. Consider Ulysses, who famously bound himself to the mast of his ship so that he could not jump overboard to join the Sirens. This technique would work in just the same way to prevent anyone else from jumping overboard, since it functions simply to deprive a person of the ability to act on any decision other than to stay where he is. Similarly, we often control ourselves by limiting the available choices. A dieter prone to temptation might do best to throw away all of the unhealthy food in his house, rather than to keep cookies around and simply will himself not to eat them.

Let us call this form of self-control "self-manipulation." This is not necessarily to say that there is anything wrong with such strategies. However, we might wonder whether it is possible to exercise self-control that is not manipulative. Can one exhibit strength of will directly, by choosing to do the right thing in the face of temptation not to, without depriving oneself of the opportunity to choose?

It will be helpful again to distinguish between the synchronic and diachronic cases. It is deeply puzzling how an agent could exercise self-control in the synchronic case, if the reason she needs self-control is to avoid doing what she most desires to do. If Anna most desires to have an assignation with Vronsky rather than to stay home, what could motivate her to exert self-control in order to stay home? On the assumption that using self-control is something that we do intentionally, it would seem that Anna would have to want most to stay home if she is going to use self-control in order to bring that outcome about. But by hypothesis, she does not want most to

stay home. We can also put the puzzle about synchronic self-control in terms of preference-satisfaction instead of motivation. On the standard way of thinking about instrumental rationality, roughly, the most rational choice is the one that has the best chance of bringing about the outcome one most prefers. But at the moment when self-control is needed, it would have to be used to bring about an outcome that the agent does not most prefer. This makes the use of synchronic self-control appear to be instrumentally irrational.

Perhaps, then, we should give up the assumption that exercising self-control is something one must do intentionally. Michael Smith and Jeanette Kennett argue, for instance, that the capacity for self-control consists in things like being disposed to think certain thoughts that help you to see the object of temptation as being in conflict with other of your important goals. For example, Esther might be disposed to start thinking of the wine as (fictitiously) poisonous rather than delicious, or to imagine her teeth turning purple after drinking it. She cannot simply choose to have such thoughts in order to bring it about that she abstains from the wine, but if she is such as to have them when experiencing temptation, this will amount to her exercising self-control in favor of abstention. And perhaps there are exercises she can intentionally perform now to bring it about that she is disposed to have such thoughts at a later moment of temptation.

A second response to the puzzle is to point out that the desire to drink the wine or have the affair are not the only relevant sources of motivation. Alfred Mele agrees with Smith and Kennett that cognitive techniques for diminishing desire are an important element of self-control, but argues that we can in fact employ such techniques intentionally and at the time of temptation. He points out that in choosing to use them, we need not be motivated only by the admittedly insufficient desire to bring about the outcome that we judge best – the desire to refrain from a second glass of wine, say. Rather, the agent might be aware of his own motivational condition and have a second, independent desire to change that condition. This second desire could be the impetus behind the intentional use of techniques to change one's own motivational structure, in a way that amounts to synchronic self-control.

At first glance, the possibility of exercising self-control over time appears less problematic. Return to the idea that instrumental rationality is a matter of maximal preference satisfaction. We assumed earlier that this should be understood as the claim that each chosen act should maximize the satisfaction of the agent's preferences *at the time* of acting. But perhaps we should instead think of rationality as a matter of satisfying our preferences *over time*. In many cases, we can foresee that we will later be motivated to make a different choice than the one we now prefer ourselves to make at that time. Anyone who has reached the point of actively trying to diet will have enough experience with temptation to be able to anticipate the siren song of ordering dessert after dinner. Such a person now prefers that he skip the sweets and will be glad after the fact that he did. This suggests that in order to best satisfy his preferences over the period of time stretching from before he goes to the restaurant until the day after his meal, he should make a choice in the morning that will prevent him from breaking his diet.

One way to accomplish this is to make a "sophisticated" choice that takes into account the prediction that if you are able to choose to order dessert at the relevant time, you will. To prevent this, the dieter could choose not to go out to dinner at all, or to bring along only enough money to

pay for an entrée. While effective, sophisticated choice can be costly; to avoid having dessert, the dieter must forfeit his dinner out or put himself at risk of having no money and suddenly needing to take a cab home. Further, it is manifestly manipulative. A less costly and less manipulative way to satisfy his preferences over time would be to make a "resolute" choice. Rather than reasoning backward from the choice he expects to make after dinner and trying to alter the options available to him then, the dieter could simply adopt a plan in the morning not to order dessert and follow through with this plan when the time comes. But again, the problem is to explain how the dieter could still be in rational control of his decision not to order dessert while going against his own preferences and evaluative rankings at that time. In other words, it still looks like an akratic decision.

A potential solution to this worry, defended by Holton, is to claim that the resolute agent simply refuses to re-open the question and make any kind of decision at the time of temptation. While it would be rationally problematic for the dieter to deliberate after dinner about whether to order dessert and choose to stick with his plan not to do so, it is rationally permissible for him to simply act on his prior resolution without deliberating. On Holton's view, weakness of the will consists in having unreasonable, over-ready tendencies to reconsider one's resolutions. Strength of will correspondingly consists in the tendency to "think less" and simply follow through with the resolutions we have formed, assuming we have not acquired significant new information in the interim.

This may well be an effective strategy for overcoming temptation in many cases. Still, we might worry that the line between being tempted to order the cheesecake and seeing it as an open question whether to do so is very delicate. The kind of temptation that is relevant to rational agency is the temptation *to do* something, and it will usually be very clear to the agent what it is and how to do it – the dieter knows exactly what he wants in that moment and how to get it. Holton therefore needs a cognitive state to exist in which the agent takes his present action to be up to him, maintains awareness of his resolution and the considerations supporting it, undergoes a shift in evaluative judgment in the light of which those considerations appear comparatively weak, and yet sees no open practical question. This seems to be a difficult state of mind to consciously maintain, bordering on bad faith. And even if it is not, it would be even more desirable to be able to choose to stick with one's resolution in light of the relevant considerations as one sees them now rather than in spite of them.

As we saw earlier, one possibility is that the agent might assign some distinctive value to following through with her plan which can shift the balance of the considerations as she sees them in favor of following through. For example, it might simply be that he places value on being a consistent, steadfast sort of person, and this is enough to overcome what would otherwise be a decisive reason to order dessert. The problem is that caring about steadfastness for its own sake can look like a mistake, since being resolute is only a good thing if it is in service of a worthy goal – otherwise it is simply pigheadedness. But whether his steadfastness is in service of a worthy goal is exactly what the agent has become confused about in the moment when resoluteness is needed. A better version of this strategy elaborates on the kind of value the agent sees in follow-through. She might see her follow-through on a particular occasion as evidence that she has the *capacity* to

be resolute when she should be, where this capacity is clearly valuable. Or she might anticipate future regret if she fails to follow through and incorporate that as a consideration into her current choice. A related idea is that following through on our plans enhances the extent to which we are self-governing agents over time, and this might be something that we care about (see Chapter Eight, section 6).

A final question concerns how seriously we should take the idea of 'willpower'. The imagery of self-control being a kind of strength, something that can be trained like a muscle and that will atrophy with underuse, is thoroughly engrained in ordinary thought and talk. We "fight" temptation, and either "overcome it" or are "vanquished" by it. To what extent is the "muscle" conception of willpower anything more than a bellicose metaphor?

The default, parsimonious position on this question is presumably that willpower is not a distinct capacity that exists over and above our beliefs, desires, and more general capacities for self-monitoring. However, a recent body of empirical data in psychology and neuroscience suggests that aspects of the strength model should be taken literally. Research done by Roy Baumeister, Diane Tice, Kathleen Vohs, and others claims to show that willpower can be depleted with use, in that subjects who have needed to use their willpower in one choice situation are more likely than fresh subjects to succumb to temptation in the next such situation. For instance, subjects who were asked to eat only radishes when chocolate chip cookies were also available, and who are then asked to complete a very difficult puzzle, give up working on the puzzle significantly earlier than subjects who were not first presented with the tempting cookies. Further, these researchers claim that performance is improved if the subjects are given glucose after the first task, lending weight to the interpretation that their capacity for strength of will had been fatigued with use and requires reinvigoration.

But while the "ego-depletion" data (as Baumeister et. al. prefer to call the phenomenon) is intriguing, other scientists have had difficulty replicating it, and still others have managed to produce the effect simply by telling subjects that ego-depletion is a real phenomenon (Dweck et. al.). As things currently stand, then, more research needs to be done before we can conclude on the basis of scientific evidence that the will behaves like a muscle in certain respects.

Summary

Akratic and weak-willed actions strike most of us as commonplace: people appear to act against their own best judgment and ignore their own past resolutions all the time. If we take this observation at face value as we begin theorizing about action, we will be led to think that the capacity to act voluntarily or intentionally must be independent of the capacity to judge what is best. It certainly seems that we can be motivated in ways that depart from our assessments of what we should do. If this is right, then accounts that tie agency too closely to rational choice or evaluative judgment must be mistaken.

That said, one person's *modus ponens* is another person's *modus tollens*. Those who are wedded to accounts of agency according to which akrasia looks impossible have been resourceful in attempting to explain away the appearance of its existence. In some cases, perhaps the agent

changes her mind about what is best at the very last minute. In others, perhaps her "judgment" about what is best is insincere; she is simply voicing the opinion of others, or of what she supposes she ought to think. In still other cases, she is actually the victim of compulsion rather than an agent who is in control of herself. The challenge is to convincingly account for all possible cases in a way that is not objectionably *ad hoc*.

If weakness of the will seems puzzling or even paradoxical, the idea of self-control is equally so. Even if an agent had the capacity to exert self-control and prevent herself from acting weakly, what could motivate her to use it? If Anna most wants to continue her affair with Vronsky, why would she exert self-control for the purpose of doing something else she wants less? And if she most wants to end the affair, then why does she need the aid of self-control to do what she most wants? Such considerations may lead us to think that the primary manifestation of self-control must be diachronic: we use it to prevent our future selves from making choices that we now disapprove of, and that we believe we will later regret. As Ulysses showed us, it can be quite effective to simply deprive one's future self of his agency, but the challenge is to explain how diachronic self-control might work without this kind of self-manipulation.

Suggested Reading

Amelie Rorty's paper "Where Does the Akratic Break Take Place" offers a nuanced depiction of the different forms akrasia can take. Donald Davidson argues that it is a failure of reasoning in "How is Weakness of the Will Possible?", while Gary Watson critiques this idea and defends his own view in "Skepticism about the Will." Alfred Mele's book *Irrationality* is a must-read on this topic, as is Richard Holton's *Willing, Wanting, Waiting*. See also Mele's exchange with Michael Smith and Jeannette Kennett ("Frog and Toad Lose Control," "Underestimating Self-Control," "Synchronic Self-Control is Always Non-Actional," and "Synchronic Self-Control Revisited"). Nomy Arpaly influentially argues that akrasia need not be irrational in "On Acting Rationally Against One's Better Judgment." And for more on the theory of weakness of will as ego-depletion, see Baumeister et. al., "Ego-Depletion: Is the Active Self a Limited Resource?"

10. Collective Agency

Up until this point, we have focused our attention entirely on individual agents and their actions, as though our topic were Robinson Crusoe alone on an island. In reality, of course, our agency is not solipsistic in this way; many of the things we do are done together with other people. We walk to class together, discuss philosophy together, and decide together what to read for next week. Athletes win games together, nations wage wars together, and citizens organize and live their lives together under a shared government and legal system. Our best theories of agency should have something to say about what happens when a group acts.

1. Questions and constraints

One major question that arises when we shift our attention to collective agency is whether we will need to introduce new metaphysical resources that are not present in our account of individual agency. Can we simply extend our best individual account to the group case, or are there features of the group case that cannot be dealt with in this metaphysically conservative way?

For instance, talk of acting together implies that there is such a thing as a group agent. We are usually happy to speak of such actions as attributable to the group and not just to its members, as when we say that the chess club organized the party. But is a group agent actually something ontologically distinct from the members of that group and the relations that hold between them? We can ask a similar question about collective intentions. We frequently assert things like "We intend to see the 8:00 showing tonight," where on the face of it, an intention is attributed to a plural subject. Is it literally possible for multiple people to "share" a single intention? Or are we implicitly talking about a set of intentions, each of which is in the mind of an individual member of the group and which together constitute a shared intention? And if it is the latter, are these just ordinary individual intentions to "do one's part," or are they an altogether distinctive attitude that we might call a "we-intention?"

A second, related question concerns the kinds of normative pressures that are characteristic of collective agency. In many cases, acting together with other people involves a set of mutual obligations. Each member of the group is accountable for doing his or her part, and the other members of the group are entitled to hold him or her to account. For example, it is normally a violation if one member of the group unilaterally drops out of participating without the consent of the others. If you and I are going to the theater together, but I peel off without warning and go to a neighboring restaurant instead, you would be entitled to object to my behavior. These observations suggest that social agency is infused with normativity. The question is whether these normative relations are instances of more general demands that we independently take ourselves to be subject to, like the demands of morality or of inviting other people to rely on you. If not,

then collective agency seems to introduce new and distinctive normative demands on us, and this should be part of our theory.

Third, there is some debate over exactly how far a successful theory of collective agency should extend. A relatively simple case involves a small group of people – perhaps only two – where no one has any special authority over the others. These people enthusiastically agree to go for a short walk together, and when the walk is over, their alliance is dissolved. But of course, many cases of doing things together with others are not so simple. Sometimes the groups are very large, as when a corporation, a university, or even a society exercises agency. It is frequently the case that there are asymmetrical authority relations within these groups; some people have more power than others over what the group does. Moreover, it is possible for some members of the group to participate in an alienated way – just for the money, say – without any real grasp or endorsement of the group's goals or the reasons behind them. Any given theory that works well for the relatively simple case may or may not extend straightforwardly to account for cases of massive, alienated, or asymmetrically structured agency.

2. Group agents

Let us start with the question of whether there are such things as group agents that cannot be reduced to sets of individuals standing in certain relations to one another. It is relatively uncontroversial that there are group agents in a weak, ontologically non-committal sense. For there to be collective action, the action in question must be attributable to something unified enough to count as the cause or source of that action. To illustrate the point, we can borrow an example from John Searle. Searle asks us to imagine that a number of people are hanging out in a public park when it begins to rain. Each of them gets up and runs to a shelter in the center of the park, and does so in a way that is sensitive to what others are doing – no one collides with anyone else – but no one's intention in escaping the rain is coordinated with or dependent on anyone else's. This is not a case of collective agency. Contrast it with a case that is indiscernible to the naked eye, but where the people in the park are in fact a *corps de ballet* performing a dance that calls for just the same movements we observed in the first case. Only in the second example are the people in the park sufficiently unified to attribute the behavior to the group as a whole, and only thus does it count as a group action.

However, what follows from this point is merely that groups must have a certain kind of structural organization. It does not necessarily follow that group agents are anything more than a collection of individuals with certain attitudes toward each other. So what considerations would lead us to countenance group agents in a more robust, irreducible sense? One suggestion is that the existence of a collective intention to act requires an irreducible group agent, or "plural subject" to which the intention can be attributed. That is, we might think that whenever a sentence of the form "We intend to A" is true, the 'we' must refer to an irreducible group agent. In part, whether or not we find this proposal appealing will depend on how we think about what it is to have a collective intention, which will be discussed in the next section. If collective intentions are themselves distinct from and irreducible to individual intentions, then this might lend support to

the idea that only irreducible group agents can have joint intentions. If they are simply individual intentions with a particular kind of content, then the proposal will be less convincing.

Even if "we-intentions" are irreducible to individual intentions, we might doubt that merely sharing one intention is enough to justify talk of a plural subject that has the intention. We ordinarily think of intentions as states or properties of a mind. But can there really be a mind that has only one mental state, or just a few? There is some reason to think that mental states can only be attributed to a subject "holistically," as part of a large network of other attitudes that hang together in a broadly intelligible way. Intentions do not normally exist in a vacuum; they are formed in light of the set of things the agent desires, as well as a number of beliefs that she has. If we think of attitudes like intention in this holistic way, attributable to a subject only as part of a larger package, then we will need more than a single, perhaps short-lived joint intention before we grant that there are such things as irreducibly plural-minded subjects.

But perhaps we do have good reason in some cases to attribute a network of attitudes to a plural subject that is not straightforwardly a function of the attitudes of the individual members. For instance, Phillip Pettit and Christian List grant that group attitudes must supervene on the attitudes of the individuals – there can be no difference at the group level without a difference at the individual level. However, they argue that there will often be no straightforward translation available from the individual attitudes to the group attitudes or vice versa. One reason is that group attitudes are "multiply realizable:" many different combinations of individual attitudes with respect to some proposition can constitute the same group attitude toward that proposition. For example, the group might count as believing that Daylight Savings Time should be abolished if all of its members have that belief, or if nearly all of them do, or if a majority of its members do. A second reason is that there can be complex organizational factors at work, such as individual specialization: the attitudes of some group members might be weightier than others in determining the group's attitudes. Perhaps only the sleep expert in the group needs to believe that Daylight Savings Time should be abolished in order for the group to believe it. And third, there can be feedback from the group level to the individual level, leading the attitudes of the individuals to evolve in light of the group attitude. Pettit and List conclude in light of these practical difficulties with reading group attitudes off of individual attitudes that "… we must think of group agents as relatively autonomous entities – agents in their own right, as it is often said, groups with minds of their own" (2011, 77-8).

More radically, we might dispense with the claim that group minds must be determined in any way by the attitudes of its members. Many philosophers of mind subscribe to some sort of "Functionalism" about mental states, according to which psychological terms such as 'belief,' 'desire,' and 'intention' refer to states of a system that play a certain functional role in that system. The role in question is given, to a first approximation, by our everyday practice of explaining and predicting behavior. Ordinary folk engage in this practice with vast success without having any real understanding of the physical architecture that instantiates this functional system; indeed, philosophers like Plato and Aristotle were able to offer deep and important insights about human psychology while thinking that cognition took place in the heart rather than the brain. And often, we can fruitfully use this practice to explain and predict the behavior of groups without any

particular knowledge of the attitudes of its individual members. We might say, for instance, that the current Supreme Court desires to promote the interests of large corporations, and that this is why it will rule in a particular way in an upcoming case.

Granted, merely being amenable to such psychological explanations and predictions is not enough to count as having genuine mentality as opposed to merely mimicking it. But since we are confident that individual human beings do have genuine mentality, we might flesh out this Functionalist schema by doing cognitive science and developing a theory of the cognitive architecture that underlies the individual case. If it is possible for a group to exhibit a relevantly similar cognitive architecture, thereby instantiating the functional roles of commonsense psychology, then we might reasonably conclude that the group itself has a mind. Whether or not this is possible, let alone likely, will depend on what the organizational structure of individual minds turns out to be. If individual minds turn out to be the product of massively distributed, modular subsystems that work relatively independently of one another, then it is less difficult to see how a group of individuals might exhibit a similar structure. If they are revealed to be more unified than this, the prospects for a collective to instantiate that structure are dimmer.

3. Collective intentions

As mentioned above, our willingness to countenance the existence of irreducible group minds or subjects might depend in part on what it is to share an intention. All parties to the debate should agree that there are such things as collective intentions in the sense that sentences of the form "We intend to A" are often true. Beyond this, of course, there is little consensus. We will examine five proposals for what collective intentions are, in what seems to me to be an organic order though an unchronological one.

a. Tuomela and Miller

According to the most minimal account, shared intention requires only that each of a set of individuals intend to take part in some group activity. More precisely, according to an account initially proposed by Raimo Tuomela and Kaarlo Miller, there is a shared intention whenever there is some group activity X such that each member intends to do his or her part of X, believes that enough of the other members will do their part, that the other preconditions of successfully X-ing are met, and that the others also have this belief. For example, you and I share an intention to go to the theater together just in case each of us intends to show up to the theater at the correct time, as part of our going together, and we each believe that this will be possible (the theater is open tonight, the subway is running, and so forth), and it is common knowledge between us that we have these beliefs. Intuitively, the idea is that shared intention differs from individual intention only in its content and associated beliefs: the former makes reference to a group activity and the contributions of others in the content of both the intention and the accompanying beliefs, whereas the latter need not. But otherwise, the structure is much the same. After all, even in the individual case, a rational agent must believe it is at least possible for her to achieve her intended action given

the circumstances she is in. The beliefs that are required in the case of shared intention simply play the same role, ensuring that each agent believes the joint action to be possible for the group to achieve together. If not enough of the others in the group do their part, then the group will fail to realize its goal.

Whether an analysis like this is fully reductive, accounting for shared intention entirely using concepts and resources that we are already committed to in our understanding of individual agency, depends on what exactly we take 'X' to stand for. It is important that each member of the collective intend to do his or her part *because* it is part of the group activity; otherwise, the strangers in the park who are individually fleeing the rain may wrongly count as acting together. After all, each intends to run to the shelter and expects the others to do so as well, which means that in some thin sense, there is a group activity "run to the shelter together" of which each intends to do what is in fact his or her part. But no individual intends to run to the shelter *because* it is part of this group activity. To exclude this kind of case, we need a more robust understanding of 'group activity' than this. On the other hand, if that understanding is so robust that it already presupposes the notion of collective agency or collective intentionality, then the account is non-reductive at best and circular at worst.

b. Searle

Searle concludes from this kind of worry that to avoid circularity, we must embrace a more explicitly non-reductive account of shared intention. He argues that "we-intentions" are attitudes that are distinct from and irreducible to individual "I-intentions," even if the latter are supplemented with extra group-related beliefs. Importantly, he does not think this means that we-intentions exist in an irreducible group mind. As Searle conceives of them, we-intentions are in the heads of the individual members of a group. That said, they are a different kind of attitude from I-intentions, just as beliefs and desires are different kinds of attitudes from either variety of intention. The difference lies in the fact that the idea of collective intentionality is built into the attitude itself (as opposed to the content of the attitude). To we-intend to do some activity X already presupposes that X will be done collectively in the robust, cooperative sense. On the other hand, the way in which each individual must do his or her part in order to accomplish the collective goal will be represented in the content of their respective I-intentions. So if Jen and Jason we-intend to make dinner – where this entails making dinner cooperatively and not just in a coordinated manner – Jen might have the I-intention "Make the ceviche as part of our making dinner together" and Jason might have the I-intention "Make the enchiladas as part of our making dinner together."

c. Bratman

Searle's brand of non-reductive account succeeds, if it does, by adding a new attitude to our model of the psychology of agency. In the interests of parsimony, we might prefer a view of shared intention that can be constructed entirely out of resources that are already present in our

best theory of individual agency. Michael Bratman attempts to offer such an account, adding complexity to the minimalist view with which we started while avoiding the commitment to *sui generis* we-intentions. His understanding of shared intention builds on the planning theory of individual intention discussed in Chapter Five. Like the Tuomela and Miller approach, Bratman holds that the primary component of a shared intention is an individual intention to do something together: "We intend to make dinner" is a matter of each member of the group having the individual intention "that we make dinner," which each person could express as "I intend that we make dinner." Against Searle's challenge, Bratman denies that the 'we' in the content of the individual intentions presupposes the idea of collective agency in a circular way. Rather, it need only refer to a group in the thin, neutral sense that includes the group of strangers in the park that are not actually cooperating with each other.

To get from mere coordination to genuine cooperation, Bratman introduces the idea of "meshing sub-plans." It is not enough on his view that each person intends to do his or her part in making dinner; in addition, each person must intend to make dinner *by way of* the others' intentions to do so. This does not simply mean that Jen believes that Jason will do his part and vice versa. It means that Jen's plans for making dinner must make reference to Jason's plans and aim to be compatible with them. Perhaps Jen intends that she and Jason make dinner together by way of her intention to make the ceviche and Jason's intention to make the enchiladas. These subplans must be non-coercively responsive to one another in ways that ensure consistency and coherence. The meshing requirement fails, for instance, if Jason blackmails Jen into making ceviche, or if Jen insists on using up all the cilantro when she knows that Jason needs some for the enchiladas. They must also be disposed not to interfere with or thwart one another, and to help each other in at least minimal ways if needed. In contrast, the strangers in the park do not intend to run to the gazebo by way of the others' intentions, and do not aim to have subplans that mesh with or positively support the plans of the others.

Further, like Tuomela and Miller, Bratman includes a "common knowledge" condition on shared intention: each participant must know that the other conditions of shared intention hold, and must know that the others know. Taken all together, these are the conditions for shared intention on Bratman's view:

> (i) intentions on the part of each in favor of the joint activity,
>
> (ii) intentions on the part of each in favor of the joint activity by way of the intentions of each in (i) and by way of relevant mutual responsiveness in sub-intention and action,
>
> (iii) intentions on the part of each in favor of the joint activity by way of meshing sub-plans of the intentions of each in (i),
>
> (iv) beliefs of each that, if the intentions of each in (i) persist, the participants will perform the joint activity by way of those intentions and relevant mutual responsiveness in sub-intention and action,
>
> (v) beliefs of each that the intentions of each in (i) are persistence interdependent
>
> (vi) the intentions of each in (i) are persistence interdependent, and

> (vii) common knowledge of (i)–(vii).
>
> (viii) the connection between the shared intention (as in (i)–(vii)) and the joint action involves public mutual responsiveness in sub-intention and action that tracks the end intended by each of the joint activity by way of the intentions of each (in (i)) in favor of that joint activity (Bratman 2014, 85-6).

To be clear, the claim is not that these conditions are necessary for shared intention, but only that they are sufficient. Though complicated in its technical structure, the core idea is relatively simple: just as individual intention should be understood by reference to its role in planning, to share an intention with other people is to plan together with them. To see what role the conditions of "persistence interdependence" are playing, let us first examine the motivations for yet another account of shared intention.

d. Velleman

J. David Velleman argues that none of the accounts we have canvassed so far truly vindicate the idea that an intention can literally be shared between multiple agents. After all, these accounts all construct shared intentions out of a set of more atomic intentions (whether I-intentions or we-intentions), such that there is really no single intention that all participants can be said to have. The problem with this is not only that we might want to hew as closely as possible to the way that we commonly speak about group agency. Velleman argues that such accounts also fail to explain how it is possible for a shared intention rationally to come into existence. Intentions are attitudes that settle things, both in the sense that they settle the deliberative question of what is to be done and in the sense that they settle what will happen. This is one of the most important ways in which intentions differ from mere desires. But how can any particular member of a group settle what the whole group will do? As Velleman puts the challenge:

> How can I frame the intention that "we" are going to act, if I simultaneously regard the matter as being partly up to you? And how can I continue to regard the matter as partly up to you, if I have already decided that we really *are* going to act? The model seems to require the exercise of more discretion than there is to go around (1997, 35).

Velleman's solution is to reject the idea that attitudes like intention must be psychological states (though of course they often are). He argues that public representations like oral speech acts and written documents can also play the settling role of intention. To serve this function, the representation must be in a position to cause the represented action to occur, and it must represent itself as playing this causal role. For instance, Jen might say aloud to Jason "I'll make dinner if you will," and Jason might reply "Sure, I will." Together, these utterances can combine to settle that they will make dinner together, since they are poised to play a causal role in bringing that action about. Given that Jen and Jason are both motivated to make dinner together, conditional on the other person agreeing, their verbal exchange can suffice to determine the matter. Importantly, like

Bratman, Velleman claims only to have articulated sufficient conditions for sharing an intention and not necessary ones.

In response, Bratman agrees that shared intentions should serve to settle whether the group performs the intended action. He denies, however, that this is impossible if shared intentions are constructed solely out of individual intentions and beliefs. What is needed instead is "persistence interdependence:" each member of the group will continue to intend the joint action if and only if the others do. Jen will continue to intend to make dinner with Jason if and only if Jason continues to intend to make dinner with her, where this is common knowledge between them. Further, Bratman adds that the group members should expect their intentions to be effective, such that they will succeed in making dinner by way of those intentions. These additional conditions are meant to allow each group member to see their own intention as settling the matter of what the group will do, even as they simultaneously see the others' intentions as settling it. Jen sees her intention as settling that they will make dinner because Jason's intention to do so is dependent on hers, just as hers is dependent on his, and because together their intentions will bring it about that they make dinner.

e. Gilbert

Finally, Margaret Gilbert argues that none of these accounts is sufficient to capture what is distinctively normative about collective agency. She proposes instead that we must understand shared intention and collective action in terms of a *sui generis* notion of *joint commitment*. According to Gilbert, instances of collective action necessarily involve distinctive mutual obligations that hold between members of the group. Each member owes his or her conformity with the group activity to the others and has the standing to demand conformity from the others and rebuke the lack of it. Further, these obligations are not escapable unilaterally, by simply deciding on one's own not to participate anymore. The other members of the group must agree to rescind the commitment to act together. Suppose Jen and Jason are making dinner together when Jason simply wanders off to the living room and starts watching TV instead. Or perhaps Jen purposefully makes only enough ceviche for one person. According to Gilbert, these actions would violate the distinctive obligations that are characteristic of collective agency. Of course, they are also *rude*, but Gilbert argues that the obligations here are not simply instances of more general moral demands on us, and that they must therefore be grounded in the nature of joint activity itself. Abraham Roth has labeled such obligations "contralateral commitments" to help distinguish them from other kinds of normative pressures.

Like Velleman, Gilbert proposes that we understand shared intention as a single commitment that involves two or more people. On her view, when multiple people commit themselves in this way, they form a "plural subject." However, she does not take this notion to entail the existence of group minds in the robust sense considered in section 1. What is distinctive about her view is that joint commitments necessarily bring mutual obligations into existence and cannot be rescinded without the concurrence of the group as a whole. These obligations are fundamentally different from the normative upshots of singular commitments to act, and so joint

commitments cannot simply be understood in terms of singular commitments on this approach. Hence, this is not an account that makes do only with the resources that we are already committed to in understanding individual agency.

Whether or not we agree that these additional resources are needed depends on the plausibility of the claim that collective action essentially involves a distinctive kind of obligation. Skeptics will attempt to account for the intuitions behind the claim by arguing that when there are such obligations, they are in fact instances of more general normative demands. For instance, many cases of collective action will involve the members of the group encouraging each other to rely on them and inducing the expectation that they can be relied on. It is plausible that morality requires us not to encourage such reliance and then intentionally fail to uphold it. Many cases also involve other kinds of personal relationships like friendship between the group members which bring their own set of demands to bear. A second strategy is to deny that all cases of collective action are such that unilateral withdrawal is a violation. In certain "no strings attached" contexts, it might be perfectly acceptable to exit a group action without permission from the others. In a nightclub, perhaps, one person might dance up to another and dance together with them for a period of time, and then dance away again without fanfare. While dancing together, the two might satisfy the conditions for sharing an intention without incurring any distinctive normative obligations to each other.

4. Acting together

Nearly all the literature in this area focuses on the nature of shared intention, out of the conviction that shared intentional action consists in the proper functioning of a shared intention. The conjecture is that when a shared intention guides and explains a group's activity, the members are intentionally acting together. To put it another way, these views hold that the difference between collective action and a mere aggregate of individual activity is the operation of a shared intention. One thing to note about this approach is that it would seem to neglect the possibility of unintended or unintentional collective action. To the extent that such things are possible, some further account will be needed that does not make essential appeal to shared intention. The primary focus of work in this area is on intentional shared action, however.

Insofar as we have an independent grip on what intentional collective action is, one way to evaluate the proposed accounts of shared intention is to consider whether they have the right scope to capture the relevant cases. In particular, all of the views we discussed are explicitly concerned not to be too inclusive; they aim not to categorize cases as collective action that are really mere aggregates of individual activity. But might they err on the other side, by being too demanding?

We should be careful to remember that some views, like those of Bratman and Velleman, explicitly claim to provide only sufficient conditions for collective action, not necessary and sufficient conditions. This means that strictly speaking, they are not vulnerable to counterexamples in which there is genuinely shared agency without satisfying the proposed conditions on sharing an intention. That said, we might prefer an account to the extent that it is less narrow in its scope, or that can be straightforwardly modified to include genuine cases that are initially ruled out.

On this measure, Velleman's proposal faces a challenge, since it seems possible for group agency to occur without any explicit verbal or written commitment that could constitute a shared intention. In some cases, a glance is enough to get started. Further, both Velleman and Bratman restrict their accounts to relatively small groups of adults with completely symmetric authority relations, such that no member of the group can simply give orders or decide for the rest. As Velleman points out, there is something especially philosophically puzzling about such cases – how we can settle on what to do together – that does not arise in the same way when one person is in a position to decide for the group. But on the other hand, a large number of the groups we might be interested in have a far less tidy structure than this. Faculties, juries, corporations, and political societies all engage in a form of collective agency, and yet they often involve hierarchies of authority and fluctuating members. It is an open question whether these accounts of small-scale collective agency can be extended to these large-scale cases that centrally involve asymmetrical authority relations.

Moreover, especially in the cases of large corporations and political societies, some of the participants might be alienated from and uncommitted to the overarching goals of the group. The employees who work in the warehouses of Amazon may not care at all whether the company succeeds in its aims of increasing market share and driving other retail stores out of business; they might simply want to earn their pay. They might also be indisposed to facilitate the efforts of the other employees, and their intention to do their part might not be dependent on the persistence of anyone else's intention to do theirs.

This is not only a challenge for Bratman and Velleman, but also for Gilbert. Gilbert argues that her "plural subject" account of shared agency is bolstered by its ability to explain the nature and existence of political obligation. Many have been persuaded by the idea that membership in a political society is enough to obligate one to support the political institutions of that society. Gilbert argues that her account is well situated to explain this fact, since to be a member of a society is to be jointly committed as a body to a certain goal. And on her view, to be jointly committed is in part to be subject to distinctive mutual obligations that are not moral in nature. In the case of political membership, she suggests that these mutual obligations include the obligation to support the political institutions of that society. This is an exciting upshot for a theory of collective agency to have. However, we might worry that the relevant political and social obligations apply even to members who are alienated from and uncommitted to the goals of the society. If so, then the obligations cannot be explained by their joint commitment to those goals.

Summary

The central question here has been whether acting together involves fundamentally different capacities or resources than acting individually. Does action at the group level introduce new types of agents, intentions, or normative obligations?

For each of these questions, there are those who are skeptical that an individualistic approach will suffice and those who argue that an individualistic account can be given. Skepticism about the existence of irreducible group agents stems largely from worries about the idea of a group

mind, which strikes many as metaphysically spooky. However, certain functionalist accounts of mentality might be more amenable to the idea that groups can have mental states like desire, belief, and intention.

Even if we reject the idea of intentions that exist in group minds, we might still think that collective agency requires a special kind of individual intention – a "we-intention." This view is motivated by the thought that the cooperative nature of the intended action must be built into the attitude itself if we are to distinguish between genuine collective action and a mere aggregate of individual doings. Against this thought, reductive approaches argue that the cooperative dimension can be captured by appeal to the group members' beliefs and other dispositions. If this is right, then we can account for the cooperative aspects of shared agency using only the ordinary notion of "I-intentions," together with the relevant beliefs and dispositions.

However, even if such an account succeeds on a metaphysical level, we might worry that it leaves out important normative aspects of acting together. If shared agency necessarily involves contralateral commitments that hold between group members and that cannot be unilaterally rescinded, then this is a further reason to think of shared intentions as distinct from individual intentions: we-intentions may have normative implications that I-intentions do not. Once again, the reductionist will respond by attempting to show that we can account for the normative aspects of shared agency by appeal to the more general demands of morality, friendship, and so forth, thereby avoiding the need to bring in *sui generis* notions of joint commitment and contralateral obligation.

Getting the right account of shared intention is important in this context because on most approaches, to act together is to implement a shared intention. Thus, questions arise about whether the accounts of shared intention we considered can capture all of the relevant cases. Some such accounts explicitly narrow their focus to certain kinds of cases – small groups without asymmetric authority relations. But we might ultimately be interested in using the lens of shared agency to think about corporations, social and political institutions, and complicated hierarchies. The ability of an account to extend to complicated cases like these may be seen as a point in its favor.

Suggested Reading

For more on the existence of group agents and group subjects, a good place to start is *Group Agency: The Possibility, Design, and Status of Corporate Agents* by Christian List and Philip Pettit, *The Ant Trap* by Brian Epstein, and "The Ontology of Social Agency" by Frederick Stoutland. Chapter Four of Carol Rovane's *The Bounds of Agency* argues that there can be a group with sufficient rational unity to count as a person, while Chapter Six of Michael Bratman's *Shared Agency: A Planning Theory of Acting Together* defends the possibility of group agents without group subjects.

Shared Agency is also the most recent and comprehensive statement of Bratman's views on collective action and shared intention. Raimo Tuomela and Kaarlo Miller's initial view of shared intention can be found in their paper "We-Intentions," while John Searle's non-reductive response

is first articulated in "Collective Intentions and Actions." Many of the essays in which Margaret Gilbert develops her view can be found in her collection *Sociality and Responsibility: New Essays in Plural Subject Theory*, while her book *Political Obligation: Membership, Commitment, and the Bonds of Society* makes the connection to political obligation. Velleman's proposal can be found in his paper "How to Share an Intention."

For questions about whether these views can be extended to the case of large-scale , hierarchical, and alienated agency, Scott Shapiro's paper "Massively Shared Agency." Shapiro also attempts to give an account of legal systems in terms of a modified version of Bratman's planning theory in his book *Legality*.

11. Concluding Thoughts

Though this book can scarcely be said to be brief, it is hard not to feel as though it has barely scratched the surface of an enormously complex topic. I will end by gesturing toward further questions that I think are promising avenues for future research, and by reflecting on the lessons that I think we should draw from what has already been done.

First, to finally lay my own cards on the table, it seems to me that the most progress over the last fifty years has been made by those who attempt to understand intentional human agency through a broadly causal lens. This is not to endorse the idea that we can give necessary and sufficient conditions for being an intentional action in purely causal terms; this strikes me as far too simplistic. It is merely to say that in my view, the most productive, generative models of agency have been those that are premised on the idea that intentional action is a matter of behavior that is caused and guided by certain psychological properties and structures – decisions, judgments, beliefs, desires, intentions, policies, and wholehearted commitments of love, to name only a few candidates. Non-causal approaches to action theory tend to be motivated largely by criticisms of the causal paradigm rather than by a clear and unmysterious statement of what the alternative is. And I do not see that any of them have successfully answered Davidson's Challenge (Chapter 3, section 4b).

That said, the criticisms posed by non-causal approaches are often incisive, and the best new research in action theory will take them seriously. Quite a bit of work remains to be done for those who would defend a broadly causal view that can do justice to the idea of practical knowledge, for example (Chapter Six), or to the insight that what we intend to do cannot be captured by an ordinary static proposition (Chapter Three, section 4d). Further, though it is a virtue of causal approaches that they can draw on the vast resources of the empirical sciences, they are sometimes open to the legitimate criticism of being too "third-personal." We should aim to understand agents from an external, theoretical perspective, but this must coincide to some degree with the first-personal, deliberative point of view. Science gives us access to the limits of our own self-understanding, but it is a mistake to excise the agent's perspective altogether and view agency entirely in functional terms.

Though the organization of the book may not suggest it, I think it is necessary for many purposes to view intentional agency as a matter of degree. Many of the classic debates concern the question of whether some incident was an intentional action or not, and this is an interesting and important question. But one conclusion we might draw from the wide range of views considered here is that there are actually a cluster of features we are sensitive to in thinking about agency, and different approaches emphasize some features over others. There will be a variety of cases in which an episode of behavior exhibits some but not all of the features we are interested in. Some behaviors are reason-guided and purposive but lack practical knowledge or awareness of what one is doing. Others involve practical knowledge but are done for no reason. Still others are clearly intentional but deficient as exercises of autonomy because they are akratic, wanton, or less than wholehearted. It is reasonable to think that action theory should be investigating the ideal form of agency *par*

excellence, in which all the features are present and full autonomy is achieved (if that is even possible). But it is also reasonable to be interested in minimal agency or any form in between – let's just be clear about what we are up to.

One of the best ways to make progress in action theory, I think, is to take the basic theoretical resources of views like those described here and see how well they can be applied in other areas. Can they fruitfully be put to use in artificial intelligence research, or in interfacing with neuroscientific research on skill and motor control? Can theories of collective agency on a small scale be extended to cover much more complex and chaotic social phenomena on a large scale? How might research in action theory inform our thinking about action and causation in the law, and how might thinking about the law inform action theory? What happens to the theory once you stop abstracting away from the real material and social conditions in which human agency actually takes place, replete with inequality, injustice, and oppression? Simplification is necessary to start, but philosophy should not end there.

Further, though I approve of Anscombe's observation that we need an adequate philosophy of psychology before we can do moral philosophy, perhaps the time has come. More work connecting the notion of moral responsibility to theories of intentional action and practical reasoning would be welcome, as would connections to normative ethics more generally (which is not to say that excellent work isn't already being done in these areas). Another area in which very interesting work is already underway, and where attention could continue to be focused, concerns the relationship between agency and time. Once we take seriously that almost all interesting cases of agency involve acting and planning over time, fascinating questions about personal identity, rational choice, and perseverance come into view. And a third important direction of flourishing research asks whether we exercise genuine agency over things other than our actions, such as our beliefs and emotions.

A general theme here is that I hope the future of action theory will aim at avoiding partisanship, disciplinary silos, and the oversimplification of opposing views. Though I am certain I have failed here to avoid this myself! At the least, I will be pleased if this book helps to counteract the trend of attributing the very simple, reductive version of Causalism to Davidson. The so-called "Standard Story," on which we can conceptually and ontologically reduce actions to events-with-the-right-kind-of-cause, is a view that is actually held by few though attacked by many. On the other hand, many Causalists have made regretfully little effort to understand or engage with Anscombe or the subsequent work she has inspired. Rather than sorting ourselves into line behind Davidson or Anscombe, let's proceed on the extremely plausible assumption that both are sources of great insight but also wrong about many, many things. And let's read science, but also literature. The problem of action is hard enough without limiting the resources on which we allow ourselves to draw.

Bibliography

Aguilar, Jesús. and Buckareff, Andrei (Eds.). (2010). *Causing Human Action: New Perspectives on the Causal Theory of Acting*, Cambridge MA: MIT Press.

Alvarez, Maria, and Hyman, John. (1998). "Agents and Their Actions." *Philosophy* 73 (2): 219–45.

Alvarez, Maria. (2007). "The Causalism/Anti-Causalism Debate in the Theory of Action: What It Is and Why It Matters." In *Action in Context*, edited by Anton Leist. De Gruyter.

_____. (2013). "Agency and Two-Way Powers." *Proceedings of the Aristotelian Society* 113: 101–21.

Anscombe, G. E. M. (1958). "Modern Moral Philosophy." *Philosophy* 33 (124): 1–19.

_____. (1963/2000). *Intention*. Harvard University Press.

_____. (1979). "Under a Description." *Noûs*, *13*(2), 219–233.

Aquinas, St. Thomas. *The Summa Theologiae*. Transl. Members of the Dominican Order. New York: McGraw-Hill and Blackfriars, 1963.

Aristotle. *Nicomachean Ethics*. Loeb Classical Library, transl. H. Rackham (1926)

Arpaly, Nomy. (2000). "On Acting Rationally against One's Best Judgment." *Ethics*, *110*(3), 488–513.

_____. (2006). *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton University Press.

Audi, Robert. (1973). "Intending." *The Journal of Philosophy*, *70*(13), 387–403.

Baier, Annette. (1970). "Act and Intent." *Journal of Philosophy*, 67: 648–658.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). "Ego depletion: Is the active self a limited resource?" *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). "The Strength Model of Self-Control." *Current Directions in Psychological Science*, *16*(6), 351–355.

Bishop, John (1990). *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press.

Brand, Myles. (1984). "Intending and Acting: Toward a Naturalized Action Theory." *The Journal of Philosophy*, *84*(1), 49–54.

Bratman, Michael E. (1987). *Intention, Plans, and Practical Reason*. CSLI Publications.

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

_____. (2006). "What is the Accordion Effect?" *The Journal of Ethics*, *10*(1-2), 5–19.

_____. (2007). *Structures of Agency: Essays*. Oxford University Press.

_____. (2014). *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press.

_____. (2018). *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford University Press.

Buss, Sarah and Overton, Lee, eds. (2002). *Contours of Agency: Essays on Themes from Harry Frankfurt*. MIT Press.

Buss, Sarah. (1997). "Weakness of Will." *Pacific Philosophical Quarterly* 78 (1): 13–44.

Castañeda, Hector-Neri. (1975) *Thinking and Doing*, Dordrecht: D. Reidel.

Chang, Ruth and Sylvan, Kurt (Eds.). (Forthcoming). *The Philosophy of Practical Reason.* Routledge Press.

Chisholm, Roderick M. (1976). *Person and Object: A Metaphysical Study*. Open Court.

Clarke, Randolph. (2014). *Omissions: Agency, Metaphysics, and Responsibility*. Oxford University Press.

Dancy, Jonathan. (2000). *Practical Reality*. Oxford University Press.

Dancy, Jonathan & Sandis, Constantine. (2015). *Philosophy of Action: An Anthology*. Wiley-Blackwell.

Danto, Arthur. (1965). Basic Actions. *American Philosophical Quarterly*, *2*(2), 141–148.

Davidson, David. (1963). "Actions, Reasons, and Causes." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

_____. (1967). "The Logical Form of Action Sentences." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

_____. (1969a). "How is Weakness of the Will Possible?" Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

_____. (1969b). "The Individuation of Events." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford University Press.

_____. (1970). "Mental Events." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

_____. (1971). "Agency." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

_____. (1973). "Freedom to Act." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

_____. (1974). "Psychology as Philosophy." Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford University Press.

_____. (1978). "Intending". Reprinted in *Essays on Actions and Events, 2nd Edition,* 2001. Oxford: Oxford University Press.

Davis, Wayne. (1984). "A Causal Theory of Intending." *American Philosophical Quarterly*, *21*(1), 43–54.

Dilthey, Wilhelm. (1961). *Meaning in History: W. Dilthey's Thoughts on History and Society*, ed. H.P. Rickman. London: Harper Row.

D'Oro, Giuseppina & Sandis, Constantine. (2013). *Reasons and Causes: Causalism and Non-Causalism in the Philosophy of Action*. Palgrave-Macmillan.

Donnellan, Keith and Morgenbesser, Sidney. (1963). "Knowing What I Am Doing." *The Journal of Philosophy*, *60*(14), 401–409.

Dretske, Fred. (1988). *Explaining Behavior*. Cambridge, MA: MIT Press.

Enç, Berent. (2003). *How We Act: Causes, Reasons, and Intentions*. Oxford University Press.

Enoch, David. (2006). "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action." *The Philosophical Review*, 115(2), 169–198.

Falvey, Kevin. (2000). "Knowledge in Intention." *Philosophical Studies*, *99*(1), 21–44.

Feinberg, Joel. (1965). "Action and Responsibility." Reprinted in *Doing & Deserving; Essays in the Theory of Responsibility,* 1970. Princeton: Princeton University Press.

Ferrero, Luca. (2017). "Intending, Acting, and Doing." *Philosophical Explorations* 20(sup2), 13–39.

_____. (Ed.). (Forthcoming). *The Routledge Handbook of the Philosophy of Agency*. Routledge Press.

Ford, Anton. (2011). "Action and Generality." In A. Ford, J. Hornsby, & F. Stoutland (Eds.), *Essays on Anscombe's Intention*. Harvard University Press.

_____. (2018). "The Province of Human Agency." *Noûs*, 52(3), 697–720.

Ford, Anton, Hornsby, Jennifer, and Stoutland, Fred. (eds). (2011). *Essays on Anscombe's Intention*, Cambridge, MA: Harvard University Press.

Frankfurt, Harry. (1978). "The Problem of Action." *American Philosophical Quarterly*, *15*(2), 157–162.

_____. (1988). *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

_____. (1999). *Necessity, Volition, and Love*. Cambridge University Press.

Frost, Kim. (2014). "On the Very Idea of Direction of Fit." *The Philosophical Review*, *123*(4), 429–484.

_____. (2019). "A Metaphysics for Practical Knowledge." *Canadian Journal of Philosophy*, *49*(3), 314–340.

Gilbert, Margaret. (2000). *Sociality and Responsibility: New Essays in Plural Subject Theory*. Rowman & Littlefield Publishers.

Gilbert, Margaret. (2006). *A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society*. Oxford: Oxford University Press.

Ginet, Carl (1990). *On Action*. Cambridge University Press.

Goldman, Alvin (1970). *A Theory of Human Action*. Princeton University Press.

Grice, H. P. (1973). "Intention and Uncertainty." *Proceedings of the British Academy*, *57*.

Harman, Gilbert. (1976). "Practical Reasoning." *The Review of Metaphysics*, *29*(3), 431–463.

Hampshire, Stuart. (1959). *Thought and Action*. University of Notre Dame Press.

Hare, R. M. (1963). *Freedom and Reason*. Oxford, Clarendon Press.

Holton, Richard. (2009). *Willing, Wanting, Waiting*. OUP Oxford.

Hornsby, Jennifer. (1980). *Actions*. Routledge and Kegan Paul.

_____. (1997). *Simple Mindedness: In Defense of Naive Naturalism in the Philosophy of Mind*. Harvard University Press.

_____. (2004). "Agency and Actions." In *Royal Institute of Philosophy Supplement* (pp. 1–23). Cambridge University Press.

Hursthouse, Rosalind. (1991). "Arational Actions." *The Journal of Philosophy*, *88*(2), 57–68.

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

Katsafanas, Paul. (2013). *Agency and the Foundations of Ethics: Nietzschean Constitutivism*. OUP Oxford.

Kenny, Anthony. (1963). *Action, Emotion, and Will*. Wiley-Blackwell.

Kennett, Jeannette, & Smith, Michael. (1996). "Frog and Toad Lose Control." *Analysis*, *56*(2), 63–73.

_____. (1997). "Synchronic Self-Control Is Always Non-Actional." *Analysis*, *57*(2), 123–131.

Kant, Immanuel. (1785). *A Groundwork for the Metaphysics of Morals.* Ed. and transl. Alan Wood. Yale University Press, 2002.

Knobe, Joshua. (2003a). "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63 (3), 190-194.

_____. (2003b). "Intentional Action in Folk Psychology: An Experimental Investigation." *Philosophical Psychology* 16(2), 309-325.

_____. (2006). "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology." *Philosophical Studies* 130(2): 203-231.

Korsgaard, Christine M. (2008). *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford University Press.

_____. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

_____. (2014). "The Normative Constitution of Agency." In *Rational and Social Agency: The Philosophy of Michael Bratman*. Manuel Vargas and Gideon Yaffe, eds. 190–214. Oxford University Press.

_____. (2018). *Fellow Creatures: Our Obligations to the Other Animals*. Oxford University Press.

Langton, Rae. (2004). "Intention as Faith." *Royal Institute of Philosophy Supplements*, *55*, 243–258.

Lavin, Douglas. (2013). "Must There Be Basic Action?" *Noûs*, *47*(2), 273–301.

List, Christian & Pettit, Philip (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Penguin Books.

Marcus, Eric. (2012). *Rational causation*. Harvard University Press.

Marušić, Berislav & Schwenkler, John. (2018). "Intending is Believing: A Defense of Strong Cognitivism." *Analytic Philosophy*, *59*(3), 309–340.

Mayr, Erasmus. (2011). *Understanding Human Agency*. Oxford University Press.

McCann, Hugh. (1991). "Settled Objectives and Rational Constraints." *American Philosophical Quarterly*, *28*(1), 25–36.

Melden, A. I. (1961). *Free Action*. Routledge.

Mele, Alfred. (1992a). *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

_____. (1992b). *Irrationality*. Oxford University Press USA.

_____. (1997). "Underestimating Self-Control: Kennett and Smith on Frog and Toad." *Analysis*, *57*(2), 119–123.

_____. (Ed.). (1997). *The Philosophy of Action*. Oxford University Press.

_____. (1998). "Synchronic Self-control Revisited: Frog and Toad Shape Up." *Analysis*, *58*(4), 305–310.

_____. (2003). *Motivation and Agency*. Oxford University Press.

_____. (2009). *Effective Intentions: The Power of Conscious Will*. Oxford University Press.

Moran, Richard. (2004). "Anscombe on 'Practical Knowledge.'" *Royal Institute of Philosophy Supplement*, *55*, 43–68.

Moran, Richard & Stone, Martin. (2009). "Anscombe on Expression of Intention." In C. Sandis (Ed.), *New Essays on the Explanation of Action*. Palgrave Macmillan.

Mylopoulos, Myrto and Shepherd, Joshua. (2020). "The Experience of Agency," in *The Oxford Handbook of the Philosophy of Consciousness*. Uriah Kriegel, ed. Oxford University Press, 164-87.

Nagel, Thomas. (1986). *The View From Nowhere*. Oxford University Press.

O'Brien, Lilian. (2014). *Philosophy of Action*. Palgrave MacMillan.

O'Brien, Lucy & Soteriou, Matthew. (2009). *Mental Actions*. Oxford University Press.

O'Connor, Timothy and Sandis, Constantine, eds. (2013). *A Companion to the Philosophy of Action*. Wiley-Blackwell.

O'Shaughnessy, Brian. (1973). "Trying (As the Mental "Pineal Gland")." *The Journal of Philosophy*, *70*(13), 365–386.

_____. (1980). *The Will: A Dual Aspect Theory*. Cambridge University Press.

Paul, Sarah. (2009a). "How We Know What We're Doing." *Philosophers' Imprint*, 9(11), 1–24.

_____. (2009b). "Intention, Belief, and Wishful Thinking: Setiya on 'Practical Knowledge.'" *Ethics*, 119(3), 546–557.

_____. (2010). "Deviant Formal Causation." *Journal of Ethics & Soc. Philosophy.*, *5*(3).

Payton, Jonathan. (2018). "How to Identify Negative Actions with Positive Events." *Australasian Journal of Philosophy*, *96*(1), 87–101.

Peacocke, Christopher. (1979). "Deviant Causal Chains." *Midwest Studies in Philosophy*, *4*(1), 123–155.

Plato. *Protagoras*. In *Plato: Complete Works*, ed. John Cooper, transl. Stanley Lombardo and Karen Bell. Hackett Publishing Company, 1997

Plato. *Republic*. In *Plato: Complete Works*, ed. John Cooper, transl. G.M.A. Grube and C.D.C. Reeve. Hackett Publishing Company, 1997.

Prichard, H. A. (2002). "Acting, Willing, Desiring." In *Moral Writings*. Oxford: Oxford University Press.

Queloz, Matthieu. (2018). "Davidsonian Causalism and Wittgensteinian Anti-Causalism: A Rapprochement." *Ergo: An Open Access Journal of Philosophy*, *5*, 153–172.

Ridge, Michael. (1998). "Humean Intentions." *American Philosophical Quarterly*, *35*(2), 157–178.

Rorty, Amelie. (1980). "Where Does the Akratic Break Take Place?" *Australasian Journal of Philosophy*, *58*(4), 333–346.

Roth, Abraham Sesshu. (2004). "Shared Agency and Contralateral Commitments." *The Philosophical Review*, *113*(3), 359–410.

_____. (2017). "Shared Agency." *Stanford Encyclopedia of Philosophy*.

Rovane, Carol. (1998). *The Bounds of Agency*. Princeton University Press.

Ryle, Gilbert. (1949). *The Concept of Mind*. Hutchinson & Co.

Sandis, Constantine. (2015). "One Fell Swoop: Small Red Book Historicism Before and After Davidson." *Journal of the Philosophy of History* 9: 372-392.

Sartorio, Carolina. (2010). "Omissions and Causalism" and "Comments on Clarke's "Intentional Omissions." In *Causing Human Action: New Perspectives on the Causal Theory of Acting*, 2010. Aguilar, J. and Buckareff , A. (eds). MIT Press, 115–134, 157–160.

Searle, John (1990). "Collective Intentions and Actions." In P. R. Cohen., J. Morgan, & M. Pollack (Eds.), *Intentions in Communication* (pp. 401–415). MIT Press.

Sehon, Scott. (2005). *Teleological Realism: Mind, Agency, and Explanation*. Cambridge MA: Bradford Book/MIT Press.

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

_____. (2008). "Practical Knowledge." *Ethics*, 118(3), 388–409.

_____. (2009). "Practical Knowledge Revisited." *Ethics*, 120(1), 128–137.

_____. (2010). "Sympathy for the Devil." In *Desire, Practical Reason, and the Good*. Oxford University Press.

_____. (2011). "Reasons and Causes." *European Journal for Philosophy of Science*, *19*(1), 129–157.

Shapiro, Scott. (2011). *Legality*. Harvard University Press.

_____. (2014). "Massively Shared Agency." In *Rational and Social Agency*, Gideon Yaffe and Manuel Vargas, eds. Oxford University Press.

Shpall, Samuel. (2016). "The Calendar Paradox." *Philosophical Studies*, *173*(3), 801–825.

Sinhababu, Neil. (2017). *Humean Nature: How desire explains action, thought, and feeling*. Oxford University Press.

Silverstein, Matthew. (2012). "Inescapability and Normativity." *Journal of Ethics and Social Philosophy* 6 (3): 1–27.

_____. (2016). "Teleology and Normativity." *Oxford Studies in Metaethics* 11: 214–40.

Smith, Michael. (1987). "The Humean Theory of Motivation." *Mind* 96(381), 36–61.

_____. (1998). "The Possibility of Philosophy of Action." In J. Bransen & S. Cuypers (Eds.), *Human Action, Deliberation and Causation* (pp. 17–41). Kluwer Academic Publishers.

_____. (2012). "Four Objections to the Standard Story of Action (and Four Replies)." *Philosophical Issues. A Supplement to Nous*, *22*(1), 387–401.

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

Stocker, Michael. (1979). "Desiring the Bad: An Essay in Moral Psychology." *The Journal of Philosophy*, *76*(12), 738–753.

Stoutland, Frederick. (2008). "The Ontology of Social Agency." *Analyse & Kritik*, *30*(2), 533–551.

Strawson, Galen. (2003). "Mental Ballistics or the Involuntariness of Spontaneity." *Proceedings of the Aristotelian Society*, *103*, 227–256.

Stroud, Sarah, & Tappolet, Christine. (Eds.). (2003). *Weakness of Will and Practical Irrationality*. Oxford University Press.

Tenenbaum, Sergio. (2007). *Appearances of the Good: An Essay on the Nature of Practical Reason*. Cambridge University Press.

Tenenbaum, Sergio. (Ed.). (2010). *Desire, Practical Reason, and the Good*. Oxford University Press.

Thompson, Michael. (2008). *Life and Action: Elementary Structures of Practice and Practical Thought*. Harvard University Press.

_____. (2011). "Anscombe's Intention and Practical Knowledge." In A. Ford, J. Hornsby, & F. Stoutland (Eds.), *Essays on Anscombe's Intention*. Harvard University Press.

Tuomela, Raimo, & Miller, Kaarlo (1988). "We-intentions." *Philosophical Studies*, *53*(3), 367–389.

Vargas, Manuel and Yaffe, Gideon, eds. (2014). *Rational and Social Agency: The Philosophy of Michael Bratman*. Oxford University Press.

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

_____. (1992a). "The Guise of the Good." *Noûs*, *26*(1), 3–26.

_____. (1992b). "What Happens When Someone Acts?" *Mind 101*(403), 461–481.

_____. (1997). "How To Share An Intention." *Philosophy and Phenomenological Research*, *57*(1), 29–50.

_____. (2007). "What Good is a Will?" In A. Leist & H. Baumann (Eds.), *Action in Context*. de Gruyter/Mouton.

_____. (2009). *How We Get Along*. Cambridge University Press.

_____. (2014). *Possibility of Practical Reason, 2nd Edition*. Michigan Publishing, University of Michigan Library.

von Wright, G. H. (1971). *Explanation and Understanding*. Cornell University Press.

Watson, Gary. (1975). "Free Agency." Reprinted in *Agency and Answerability: Selected Essays,* 2004. Oxford University Press.

_____. (1977). "Skepticism About Weakness of Will." Reprinted in *Agency and Answerability: Selected Essays,* 2004. Oxford University Press.

_____. (1982). "Review of J. Hornsby, *Actions*." *The Journal of Philosophy* 79(8), 464-69).

_____. (1987). "Free Action and Free Will." Reprinted in *Agency and Answerability: Selected Essays,* 2004. Oxford University Press.

_____. (2003). "The Work of the Will." Reprinted in *Agency and Answerability: Selected Essays,* 2004. Oxford University Press.

Wedgwood, Ralph. (2006). "The Normative Force of Reasoning." *Noûs*, *40*(4), 660–686.

Wegner, Daniel. (2002). *The Illusion of Conscious Will* . Cambridge, MA: MIT Press.

Wiland, Eric. (2012). *Reasons*. Bloomsbury Press.

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

Wilson, George and Shpall, Samuel (2012). "Action." *Stanford Encyclopedia of Philosophy*.

Winch, Peter. (1958). *The Idea of a Social Science and its Relation to Philosophy*. Routledge.

Wittgenstein, Ludwig. (1953/2010). *Philosophical Investigations*. John Wiley & Sons.

Wolf, Susan. (1987). "Sanity and the Metaphysics of Responsibility." In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46–62). Cambridge University Press.

Yaffe, Gideon. (2010). *Attempts: In the Philosophy of Action and the Criminal Law*. Oxford University Press.