Quaestiones Infinitae

PUBLICATIONS OF THE DEPARTMENT OF
PHILOSOPHY AND RELIGIOUS STUDIES
UTRECHT UNIVERSITY

VOLUME XCV

Cover image and design by Michiel Janssens, 2016.

ISBN 978-94-6103-055-9

Freedom & Self-Knowledge

Vrijheid en Zelfkennis

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 28 september 2016 des middags te 12.45 uur

door

Dawa Ometto

geboren op 3 november 1986

te De Bilt

Promotoren: Prof. dr. T. Müller
Prof. dr. M. Düwell
Copromotor: Dr. A. Kalis

The research leading to this doctoral thesis was funded by the Netherlands Organisation for Scientific Research (NWO), Project Number 322-20-005.

Acknowledgments

They say writing a dissertation takes three to four years, but now that it is finished it seems to me that it really takes one's entire life. I don't just mean that it feels like it took a lifetime to complete this thesis. Looking back, I now realize that the process which culminated in this book started much earlier: with the great friends and teachers I had in university and high school, and before that, with the obsessive reading of books when I was young. I therefore want to begin with those who have been with me from the start. First of all, I want to thank my mother Antonella, who always fostered my curiosity and told me I could do anything at all if it made me happy—including doing something crazy like studying philosophy. The rest of my family has always been incredibly supportive, and I thank them for everything: my grandparents Atie and Luuk, and my uncle and aunt René and Marjan. I also want to thank my cousins Hannah, Lucas, and Aster (and later her husband Joost) for their support and friendship.

Approaching more recent times, I want to thank my promotor and supervisor Thomas Müller, who first took me on as an assistant and then encouraged me to pursue a PhD. I have deeply benefited from his advice, his teaching, and his thoughtful and sharp comments on my writing. I also thank my other promotor Marcus Düwell who, like Thomas, was instrumental in helping me to secure the grant which allowed me to perform this research. This also goes for my co-promotor, Annemarie Kalis, who never tired of reading yet another version of my research proposal. In the last two years of my PhD, Annemarie has also been an amazing supervisor, who always kept my spirits up—and my feet on the ground, when necessary. Thanks for all your help in writing this thesis, for all your advice, and for keeping things fun all at the same time.

When studying at Utrecht I had the great fortune of meeting some very bright minds and great friends: Niels van Miltenburg and Jesse Mulder. As students they were already hugely influential on my philosophical upbringing, and this has only increased in the time that we have worked together as researchers in the department.

There is not a good thought in this book that I don't owe to our ongoing discussions. Thank you for continually showing me that philosophy at its best is interesting and worthwhile. We were later lucky enough to be joined by Antje Rumberg. Together, Niels, Jesse and Antje are largely responsible for dragging me through the occasional mood swings that came with writing this thesis and exploring the world of academic philosophy. Thanks for always being there for me, and for all the great laughs!

Among other academic friends whom I have had the pleasure of meeting, and the benefit of discussing my work with, are (in no particular order): Florian Fischer, Adrian Haddock, Michael De, Rebekka Gersbach, Jens Gillessen, Daan Evers, Charlotte Alderwick, Christian Kietzmann, Nicole Del Rosario, Harmen Ghijsen, Bart Kamphorst, Dascha Düring, Verena Wagner, Marius Backmann, Andries De Jong, and Daan Dronkers. I also want to thank audiences and participants at the 2014 Summer School 'Powers, Perception & Agency', the Utrecht Practical Philosophy Colloquium, the 2015 OZSW Conference, the Dutch Analytic Research Seminar, and the students of the RMA Philosophy at Utrecht. Thanks to Erasmus Mayr for his extensive comments. Special thanks go out to Biene Meijerman and Suzanne van Vliet, who were of great help with organizational matters.

I have also been lucky enough to have a good deal of friends outside of academic philosophy, who have kept me in touch with the real world, and have done more to keep me going than I think they know. It would be impossible to list them all, but I must mention Michiel, Rachelle, Janna, and Erik. Thank you for all the good times, and for believing in me almost as a matter of course. Specials thanks are also due to the staff of Café Derat, The Village, and Blackbird.

When staying as a visiting fellow in Leipzig during the summer term of 2016, I had already finished most of this dissertation. Nevertheless, I want to thank some of the philosophers I have met there for their inspirational work and teaching: Matthias Haase, Doug Lavin, Wolfram Gobsch, and Sebastian Rödl.

Lastly, I owe greatest thanks of all to Candice Cornelis, whose patience with me in the last four years has bordered on the saintly. Without her support, and without all the happy times we've celebrated together, I would be nowhere. I dedicate this work to her, with more love than I know how to express on these pages.

Dawa Ometto
Saturday August 20th, 2016

* * *

Contents

Introduction: free will, accidentality, and spontaneity

Freedom of the will is a never-ending source of puzzlement for academic philosophers. At the same time, it is something deeply familiar to everyone. For the relevant concept of freedom underlies much, if not all, of our ordinary discourse and thinking about ourselves and others. We make a clear distinction between behavior we take to be the expression of somebody's own choice, and things that are not up to that person: for example, one cannot help but shiver due to the cold, but one *can* make up one's mind to turn up the heat in the room—or not. And out of the many things we could say that someone *is doing* on a given occasion—e.g., changing the acoustic properties of the room just by sitting there, or exercising gravity on the objects around her—we usually have no problem isolating the things that belong to the latter category: her *actions*.[1] The concept of free choice thus structures our perspective on others. But, perhaps most importantly, we are familiar with that concept *as agents*. In making up our minds about how to act, we regard ourselves as settling the question 'what will I do?'. When we engage in practical deliberation about that question, we thus presuppose that the outcome depends on that deliberation, and is thus genuinely *open*. Before we make up our minds, it is not yet decided what we will do—*we* decide that. We thus regard our actions as events that did not have to happen necessarily: we *could* have refrained from performing the action we did perform. Without understanding ourselves as free in this way, we would not know how to begin to make sense of the world around us and our relation to it.

However, our attempt to make sense of the world, through science in general and philosophy in particular, can itself cast doubt on the validity of this conception of freedom. Are we not ourselves part of the world whose workings we study?

[1] Compare Anscombe [1963, p. 8].

And are we not, then, subject to the same laws of nature that we have discovered to govern that world? If so, should we not admit that our actions are simply physical events, mere links in a long chain of cause and effect? And so, should we not conclude that the special perspective we take on our own actions and those of others is perfectly illusory? For a long time, this worry was phrased almost solely in terms of the problem of *universal determinism*, which is, roughly, the view that whatever happens happens necessarily, so that the future is never in fact open and it is not true that we could ever *not* have acted as we did. In more recent times (at least since the latter half of the 20th century) this has changed in the face of widespread agreement by scientists that not everything is, in fact, determined: many believe there is fundamental indeterminism at the subatomic level. Still, scepticism about free will has remained common: at the macroscopic level of human bodily movement, it was supposed, determinism may still reign.

But even more recently, many philosophers have recognized that the question whether we can somehow reconcile the familiar conception of agency with determinism may be a red herring. Free will, they argue, may be illusory even if our actions are *un*determined. For if indeterminism is the absence of determination, what is it but pure randomness of the kind that does indeed exist at the subatomic level? If that is true, the idea that there is a special class of happenings which find their origin in human choice seems to be at least as problematic as it would be under determinism: what we regard as our choices would be nothing but random events, over which we have as little control as over the outcome of a die roll. What we end up doing would be merely *accidental*, and precisely not up to ourselves. Thus we are faced with a dilemma: we either admit that our actions are determined, or that they are random accidents. Both possibilities seem difficult to reconcile with the familiar conception of agency.

Of course, philosophers old and modern have tried to escape this dilemma, most often by arguing that free action need not be undetermined. So-called *compatibilists* give up the idea that practical deliberation is truly open-ended, and thus that we could *not* have acted as we did. Yes, they argue, our actions may be links in a giant causal chain. But some of those links just *are* the elements of our practical deliberation: our beliefs, desires, and intentions. Acting freely consists, at least in part, in being *caused* to act by those parts of our psychology.

However, many agree, the compatibilist response to the dilemma is hardly appealing. If an action is caused by some psychological state of the agent's, which is in turn caused by other psychological or physical states—and so on until we reach the beginning of time—then in what sense are we free to make up our minds about what to do? At the same time, it seems to be no good simply to insist that the causal chain

must break off somewhere: that is just to invite randomness, making the first link in the chain a mere accident. Neither possibility, it seems, accommodates the idea that in practical deliberation, *we ourselves* determine what to do on the basis of our reasons. That idea, paradoxically, seems to include *both* an element of accidentality and non-accidentality: responding to one's reasons for acting, intuitively, is not *arbitrary*, nor is it the same as being psychologically *necessitated*.

My aim in this thesis will be to show how to overcome this dichotomy between accidentality and causal necessitation. A sound account of agency, I will argue, has to take seriously the idea that in practical deliberation, we are genuinely *self-determining*. I will show that progress towards such an account has been hampered by the prevalence of a *reductive* understanding of intentional action, according to which the explanation of an action in terms of an agent's reasons is just a species of the 'ordinary' form of quasi-mechanical causation we associate with the laws of nature. On that assumption, non-accidentality can consist only in something's being caused by prior states and events. Instead, I defend a less dogmatic view, on which we can distinguish a number of formally distinct kinds of causal principles, or forms of non-accidentality. One of those is the kind of responsiveness to reasons, or *spontaneity*, that is exhibited by intentional action. In exercising such spontaneity, I will argue, we are truly self-determining and free.

The resulting account will do justice to the familiar conception of agency, including the idea that in deciding what to do, we settle upon a course of action that was not necessary before that point. So I defend a variant of the position called *libertarianism*: the view that free will must involve indeterminism, and that we actually possess free will. However, my position differs from existing libertarian views in a crucial respect. For contemporary libertarians actually share the reductionist picture of their compatibilist competitors: they accept that acting intentionally consists in being caused by psychological states. And so they agree that there could be intentional agency in a deterministic world. On their view, it is just that such deterministic-but-intentional action would not count as *free*—for that, it is required that the causal link between psychological states and action be probabilistic. They thus distinguish between acting intentionally and acting freely.

I believe this is a grave mistake: once we have accepted the reductionist picture, it becomes impossible to see what the introduction of indeterminism could add in terms of freedom. Indeed, I argue that the reductionist picture of action is best seen as an inherently compatibilist doctrine. So in contrast to almost all contemporary accounts, my view is that we must *not* separate the notions of intentional agency and freedom: acting freely just is acting for a reason.[2]

[2]Helen Steward forms an important exception to the consensus that action is compatible with deter-

In defending this picture, I face two tasks. First, I have to argue that the reductionist (and compatibilist) theory of action fails. Second, I have to explain how we can reconcile the idea that acting for a reason is a *sui generis* form of causality with the fact that actions are real-world, physical movements, and so arguably governed by the laws of nature. The reductionist tries to reconcile these two aspects by claiming that reasons-explanation just is ordinary causal explanation. In contemporary philosophy, it is all too often assumed that this reduction is inevitable—so much so that it is often forgotten that it *is* a reduction. Arguably, this is exactly because it is unclear how *else* we could reconcile spontaneity with the laws of nature. The two challenges must thus be faced together. By analyzing the nature of the rational kind of non-accidentality, I will both show how the reductionist picture fails to accommodate it, and explain how the reconciliation between spontaneity and laws of nature *can* be effected—thus undermining the very need for a reductionist account.

To better understand the challenge I will take up, it will be instructive to reflect on hitherto existing appeals to the spontaneity of practical reasoning and action.[3] A philosopher who has famously resisted reductionism, and has done much to recover an understanding of the spontaneity of reason, is John McDowell [1994]. Yet as his work also illustrates, we are still far from an answer to the question how rational spontaneity can be accommodated in the natural world. Let us briefly consider why.

According to McDowell, there is a strict distinction between what he calls the space of reasons and the realm of law: respectively, the 'logical space that is the home of the idea of spontaneity' [McDowell 1994, p. 74], and that 'within which natural-scientific description situates things' [McDowell 1994, p. xix].[4] Now as McDowell recognizes, such a strict separation risks making it unintelligible how spontaneity can be at work in nature. And so it risks making it unintelligible how our actions could be both spontaneous *and* part of nature. McDowell's proposed solution is to resist the underlying dogma that nature is exhausted by the realm of law. Instead, we should

minism [Steward 2012a]. She therefore calls her position *agency-incompatibilism*. Although I would not object to calling my own position a variant of agency-incompatibilism, there are important differences between my position and Steward's. I explain these differences at length in §5.3. In short, they come down to the following. First, unlike Steward, I take determinism to be incompatible only with specifically *intentional* or *self-conscious* action. Second, I will argue that Steward's position does not (where mine does) exclude what I will call 'higher level' determinism. I thus believe my position is unique within the scope of the contemporary analytic debate.

[3] This appeal is not new or modern: it is familiar from a philosophical tradition which opposes the empiricist and reductionist tradition which has so heavily inspired current analytic philosophy. The arch-proponent of the distinction between rationality or spontaneity and law is, of course, Kant. But there are many others: obviously, Kant's successors in the idealist tradition, but we arguably also find the same idea in Wittgenstein and the Aristotelean turn given to his philosophy by his disciple Anscombe. I will explore some important parallels and differences between these different ways of thinking about spontaneity in chapter 5.

[4] McDowell's primary focus in *Mind & World* is on spontaneity as it relates to perception, and not action. Nevertheless, the problems we face in the theoretical and the practical domain seem structurally similar, as McDowell himself notes [1994, p. xxiii].

also recognize *second* nature, the set of rational capacities that 'belong to a normal human organism' [McDowell 1994, p. 84] and which is developed by initiation into cultural and linguistic practices. Undoubtedly, that is a step in the right direction. However, it seems that it is not enough. For, as McDowell admits, second nature must have a 'foothold in the realm of law' [McDowell 1994, p. 84]: the capacities for theoretical and practical reasoning are capacities of beings—ourselves—that also have a *first* nature. But then we must ask: if we are, *qua* first-natural beings, governed by 'blind' laws, how is it possible for spontaneity to be at work in the actual movements we make?

Thus I believe it is not enough to insist on the idea that reason and law belong to two distinct 'logical spaces'. For then we cannot escape the conclusion that we, human agents, belong to *both*, without being able to understand how this is possible. So, instead of restricting reason to its own logical realm, I suggest that we must develop a picture on which reason can genuinely be seen to be the governing principle of real-world movements—without reducing it to blind causality. At the same time, we must of course stay clear of quasi-solutions to this problem such as dualism. McDowell is aware of this challenge. He says that if 'everything in the world happens according to laws of nature' [McDowell 1999, p. 102], a thesis derived from Kant, it will be difficult to reconcile this with the idea that there is true spontaneity:

> If the occurrences that are [. . .] our actions are all, under some description, displayable as cases of the operation of law, how can they count as free simply on the grounds that they are also susceptible to other descriptions [as intentional actions, D.O.] under which they are not subsumable under law?[5] If one is impressed by such questions, the simplest move is to reject the Kantian thesis, saying that only some of what happens in the world is subsumable under natural law. But if we make that move, we do seem to need to say more about how the law governed and the free are related, especially given how plausible it is that natural law holds sway at least over the sub-personal machinery that underlies our ability to act and think. [McDowell 1999, p. 102]

That is the challenge I think we must face. We must explain how an action can be self-determined, free, and spontaneous, while not violating the laws that 'hold sway [. . .] over the sub-personal machinery that underlies our ability to act' [McDowell 1999, p. 102]. The answer, I think, lies partly in embracing indeterminism. If our actions are at the same time *both* acts of spontaneity *and* physical movements in the natural world, then spontaneity cannot be allowed to *infringe* on the laws of nature. Instead, the laws of nature *must leave room* for our practical reasoning to have influence on the course of events: they must leave open multiple really possible ways for the future

[5] McDowell is here criticizing Davidson's 'domestication of Kant' [McDowell 1999, p. 102] in [Davidson 1970].

to develop.

This strategy has been overlooked by those who agree that reasoning cannot be reduced to the operation of blind causes. This is, I think, because it is all too often tacitly assumed that 'realm of law' must mean 'realm of determinism'.[6] And when it is pointed out that this is a mistake, the response is often that this is irrelevant for the reasons we have seen above: what can the *absence* of determination by physical law do to render an action free? It is thus overlooked that the absence of physical determinism may be a necessary, though not sufficient, condition of freedom. It may be that, if reasoning constitutes a *sui generis* principle of movement, this entails that there must be movements that are not *fully* determined by other causal principles, i.e., by laws of blind causality. That is the picture I will defend.

Of course, embracing indeterminism is not enough: we also want to understand *how*, or in what way, practical reasoning is a causal principle. Here we should be careful about what to expect. Since the idea is that reasoning is a *sui generis* form of causality, we should not expect an account that shows how reasoning leads to action in the *same* sense as one might give an account showing how a button-press on a machine leads to the dispensing of a can of soda. If I am correct in resisting reductionist, causal theories of intentional action, such an account simply does not exist. But still, we want to know how something that appears to belong to the space of reasons can be a principle of change in the natural world.

That is where the second topic in the title of this dissertation comes in: self-knowledge. Although the term self-knowledge has many applications, I will use it to mean the non-observational, first-personal knowledge we have of our own thoughts. On first sight it may be difficult to see what such knowledge could have to do with the challenge I have identified. But according to Elizabeth Anscombe, there is actually a very close link between agency and self-knowledge. According to her account of action, which I will defend as an alternative to the reductive orthodoxy, intentional action is essentially characterized by the agent's knowledge of what she is doing: her *practical knowledge*. Such knowledge, as we will see, is spontaneous: it is not acquired by *observing* what one is doing, but by reasoning about what to do. Yet at the same time, it is knowledge of an action, and that is, a real-world *movement*. In practical knowledge, then, 'self-knowledge [. . .] extends beyond the inner recesses of the mind, beyond the narrowly psychical, and into the things that I am doing' [Thompson 2011, p. 200].

I will argue that by understanding how self-knowledge can do this—i.e., extend beyond the psychical and into the real-world movement that is one's intentional

[6] This is certainly the case in Kant's formulation of the third antinomy, where he is clear that for something to happen according to natural law is for it to have a sufficient prior cause [Kant 1999c, A474].

action—we can understand how spontaneity can be at work in the natural world. By understanding practical knowledge, we can thus understand the kind of causal principle that is practical reasoning. As we will see, the relevant form of causality is essentially self-conscious: it operates through the agent's knowledge of the causal relation between her reasons and her action. I will explain what this slogan—reasoning is essentially self-conscious—means in due course. But if I am right, it follows from this that practical reasoning cannot be reduced to the 'ordinary' causality which belongs to McDowell's realm of law. Hence, I will argue that acting intentionally, on the Anscombean picture, is self-determination, and hence physically undetermined.[7] If that is right, the familiar understanding of ourselves as agents will prove to be more than an optional conception, to be dismissed as mistaken by compatibilists or skeptics about free will. Instead, that understanding will be the very thing our freedom consists in: practical self-knowledge.

Method

Writing a method section for a work of philosophy is notoriously difficult. In such sections we often find the author claiming that he or she will engage in 'conceptual analysis'—a term which has become almost synonymous with doing analytic philosophy, but of which the meaning remains hotly disputed. Why is it so difficult to say plainly what the method of (analytic) philosophy is? The answer to this meta-methodological question, I think, is the following. In philosophy, method and substantive inquiry just are not cleanly separable. This is because one of the main *objects* of philosophical investigation is at the same time that by which we conduct that investigation: *thought*. Understanding conceptual analysis as a method is, in part, to understand what *concepts* are, and how they relate to reality. Thus one's answer to the philosophical question 'what is thinking?' will significantly constrain the range of possible answers to the methodological question, and vice versa.

If that is right, then the development of a philosophical methodology and the defense of a substantive philosophical picture have to go hand in hand. And, as I will now try to explain, this seems especially true in the case of the main subject of this essay: intentional and free agency. In developing an account of this phenomenon, I take myself to be articulating the contents of our self-understanding as agents, along the lines sketched at the beginning of this introduction. However, in doing so we face a peculiar challenge. For the concept of agency seems to *depend* on our

[7] The Anscombean picture of intentional action is currently undergoing an active revival. See, for example, Rödl [2007], Thompson [2008], Lavin [2015]. However, it has not yet been explicitly connected to the question of free will and determinism. By doing so, I thus hope to contribute to both our understanding of free will, and the emerging Anscombe-inspired theory of action.

self-understanding in a way that, say, the concepts of mass and water do not. It is difficult to spell out the difference in detail. But clearly, the latter concepts are concepts of things that are what they are independently from what we *take* them to be. And that does not seem to hold for the concept of agency. For a movement to be an intentional action, and not just a bodily movement, the agent's ability to understand the movement *as* an action seems to be somehow essential. If the agent were wholly unable to understand the movement as a contribution to some goal of hers, we could hardly insist against her that it *is* in fact something she is doing intentionally.

Without our ability to understand ourselves as agents, then, it seems there would not *be* such a thing as agency in the first place. The concept of an action is the concept of a movement to which we are cognitively related in a distinctive way. Methodologically, this has the following consequences: a sound account of agency must at the same time be an account of how we come to understand ourselves as agents. That is, the question 'what do we think ourselves to *be* when we think of ourselves as agent?'—what the *content* of our self-understanding is—cannot be isolated from the question of what it is to *think* of ourselves as agents, i.e., of what is it to *understand* ourselves as such. This does not mean that it is impossible to find out what agency is. Rather, it shows that the answers to our two questions (what *is* agency, and how can we determine what agency is?) must necessarily be developed in tandem.

There will be some who see this methodological complication as a reason to be skeptical about philosophy's ability to acquire genuine knowledge (at least when it comes to our topic). But such scepticism is unwarranted. It dogmatically assumes that our self-understanding cannot amount to knowledge—perhaps on the basis of the intuition that knowledge is only possible when it comes to *empirical* concepts, such as water and mass, which do not depend on our point of view in the peculiar fashion I have sketched. Yet it is important to see that this is not a neutral theoretical starting point, but a positive commitment to the thesis that there can be no such thing as *self-knowledge*: non-observational knowledge that is *grounded* in the peculiar relation between the thinking subject and the object of her thinking—in case of our topic, her intentional action.

As we have seen, it is part of my aim to show that such knowledge *is* possible. And if I am right, understanding such knowledge simultaneously is to understand its object—intentional action in the case of practical self-knowledge, and belief in the case of theoretical self-knowledge. For self-knowledge, on my view, is the conclusion of (practical or theoretical) reasoning. And so an account of how we *know* what action is will at the same time be an account of action. For my topic, it is thus necessary that

the development of a substantive account and methodology go hand in hand.

Because of this approach to free will and self-knowledge, the questions I will focus on differ from those that are typically asked in the contemporary free will debate. In that debate, the question what it is to act freely is often approached as follows. It is noted that it is a platitude that, if someone does not act freely, she will not be morally responsible for what she does. So, it is assumed, we can test a theory of free will by seeing whether we would be inclined to hold agents morally responsible in certain circumstances. For example, some will try to show that we would not hold someone responsible if we thought her actions were physically determined—and others will try to show the same for the case of indeterminism. Such arguments commonly work by analogy: would we be inclined to hold someone responsible if they were coerced by an evil neuro-scientist? If not, then why should we hold them responsible under determinism? I think this method has some serious drawbacks. For one thing, it is not at all obvious that we have a clearer grasp of the concept of moral responsibility than of the concept of free will.[8] Also, the method has led to the proliferation of ever more fanciful examples and counterexamples. These often obscure important questions. For example, does it even make sense to say that someone is forced to *act* by a device implanted in her brain? Is acting for a reason compatible with (in)determinism? By focusing exclusively on moral responsibility, we will never learn an answer to these questions.

Since I want to argue that intentional action is intrinsically free action, I must thus reject reflection on the connection between free will and moral responsibility as a starting point. That does not mean that I think free will is entirely unrelated or unimportant to ethics and moral responsibility: on the contrary. Nor do I believe that the importance of free will is exhausted by its implications for ethics and moral responsibility: it is also important because it forms such an integral part of our self-understanding. Although understanding the relation between freedom and morality falls beyond the scope of this thesis, I also believe that doing so requires that we first have a sound account of the self-determining, spontaneous character of human agency. And so we cannot make use of intuitions about morality and responsibility in developing that account. Moreover, it seems that the exclusive focus on moral responsibility is itself a symptom of the fact that the contemporary debate does not thematize the tension between determinism, indeterminism, and our self-understanding as agents.

[8] This becomes clear when we notice that one major position in the debate about moral responsibility is so-called semi-compatibilism, which claims that moral responsibility, but not free will, is compatible with determinism [Fischer and Ravizza 2000]. Semi-compatibilists thus reject even the putative platitude that we cannot be responsible without being free. In the face of this, it becomes hard to see how the notion of moral responsibility could provide us with any constraints in thinking about free will.

As I explained, my approach to the topic of free will requires a broad understanding of how human agency can fit into the natural world. This means that this thesis cannot be *just* about free will: inevitably, we will encounter questions belonging to other specialized philosophical disciplines: as we have already seen, questions about intentional action and self-knowledge, but also about, e.g., *causality*, the *mind*, and even *life* itself. Part of what I want to show is that, as long as we cling to a broadly empiricist and reductionist understanding of these topics, we simply lack the resources to arrive at a sound account of human freedom. Therefore, in the course of my investigation, I will defend alternative, anti-reductionist accounts of the topics I mentioned. Many of these accounts are currently being developed outside of the context of the free will debate. Each of these topics deserves a book-length treatment of its own, and I am aware that I cannot hope to definitely settle the respective debates about them in the space of this thesis. So although I will attempt to show why, e.g., a non-reductive powers-based account of causality is convincing in its own right, my primary purpose will be to show how the various elements of the anti-reductionist picture can be combined into an account of freedom that does justice to our self-understanding. If successful, this in itself constitutes a strong argument in their favor.

The reader will notice that in resisting the mainstream reductive treatment of many of the topics I mentioned above, I will often appeal to the work of G.E.M. Anscombe (and philosophers who are inspired by her approach). In the first instance, this is because her philosophy contains many unjustly ignored insights of great importance to my aim of developing a sound account of free action. But I also hope to show that Anscombe's writings on, e.g., intentional action, the first person, and causality form a coherent whole. It is only by attending to this unity that we can prevent misunderstandings of her philosophy. For example, I will show (in chapter 4) how the failure to take seriously her account of first-person thought has lead to reductionist interpretations of her account of practical knowledge and intentional action. Now, Anscombe herself (or more precisely, her Wittgensteinian heritage) is undoubtedly partly responsible for the fact that the systematic connections between the different aspects of her work are so often ignored. But it would be wrong to insist that there cannot be such systematic relations because she is, after all, a Wittgensteinian. For I believe the interest of Anscombe's philosophy lies precisely in the fact that she herself saw no difficulty in combining Wittgenstein's teachings with a form of analytic (and systematic) Aristoteleanism. By focusing on the latter aspect of her work, we can learn what it could be for analytic philosophy to be systematic, but not reductive.

Structure of the dissertation

In **chapter 1**, I will begin by providing an overview of positions in the contemporary debate about free will. I focus especially on the so-called *luck objection* to libertarianism: the argument, or family of arguments, that tries to show that undetermined actions are indeed lucky, or accidental, in a sense that undermines their freedom. Discussion of the luck objection has increased in recent years, and it is in that discussion that we find the best contemporary rendition of the idea that free action must be non-accidental. The various formulations of the objection all try to show that, if an action were undetermined, an agent could not have *control* over it. I review several attempts of libertarians to counter the objection, and argue that they fail.

Interestingly, however, there does appear to be a cogent response to the objection: namely, to insist on a certain *reductive* account of agential control. If one accepts that account—the Causal Theory of Action (CTA)—such control consists in the causation of the action by certain of the agent's psychological states. On that theory, some so-called *event-causal* libertarians have recently argued, it is possible to have control over an undetermined action, namely, when the causal relation is probabilistic. I argue that this response comes at a huge price for the libertarian: while explaining the *possibility* of undetermined free action, it makes it impossible to insist that indeterminism is *necessary*. For if agential control just consists in causation by psychological states, then it is perfectly compatible with determinism as well. I conclude that all currently existing formulations of libertarianism are flawed. They try to introduce the requirement of indeterminism after already buying into a reductive picture of agency whose very purpose it is to make compatibilism possible. This has happened because participants in the debate fail to notice a crucial assumption underlying the contemporary discussion: that free action is intentional action which satisfies certain further conditions, rather than intentional action *simpliciter*. I thus suggest that any sound libertarian theory of free will must begin by rejecting the distinction between free action and intentional action. And that means that she must reject CTA in favor of an inherently incompatibilist theory of action: a theory on which acting for a reason *as such* requires indeterminism. At the same time, of course, such a theory must avoid the luck objection.

However, it is not yet clear how these two requirements can be combined. What could it mean to say that an action is undetermined, not caused (as on the event-causal model), yet not a matter of luck? In **chapter 2**, I set out to investigate this. I argue that our puzzlement about this question is grounded in a certain picture of what it means for something to be a matter of luck (§2.1). According to this picture, which I call the *uniform conception of luck*, an event is (approximately) lucky if it does not occur in a

wide range of close possible worlds. To show this, I compare the discussion of the luck objection with another philosophical debate in which luck plays an important role: that on *epistemic luck*. I argue that the uniform conception of luck fails to achieve its purpose: it cannot explain why accidentality is incompatible with knowledge and freedom, respectively. Instead, I propose that we should look at non-accidentality in a different way: for something to be non-accidental is for there to be an *explanation* of why it happens (§2.1.2). And, I argue, for an action not to be accidental in the sense that threatens freedom is just for it to have a reasons-explanation, i.e., to be intentional.

Now, CTA offers us a reductive account of reasons-explanation: reasons are causes of the same kind that we find everywhere else in nature, and thus event-causation is the *only* kind of non-accidentality. Once we accept this reduction, compatibilism becomes inescapable. But, I argue (§2.2.1), there is no need to accept this picture. Proponents of CTA are correct to say that reasons must be causes of some kind—they must make it no accident that an agent acts as she does. But it does not follow that we must accept the reduction of reasons-explanation to ordinary causal explanation. Moreover, I suggest that there is reason to suspect that CTA's attempt to reduce the non-accidentality inherent in intentional action to event-causation fails (§2.2.2). As I explain, that is arguably the lesson to draw from the well-known problem of deviant causal chains. Although I will not yet try to refute CTA definitively (I defer that until chapter 4), I argue that there is enough reason to suppose that a sound account of freedom requires us to develop a non-reductive account of intentional action and (therefore) of practical reasoning.

In the following two chapters (3 and 4), I thus work on the assumption that reasons-explanation is a form of non-accidentality irreducible to ordinary event-causation. By developing such a non-reductive account, I will at the same time attempt to uncover the fatal flaws underlying CTA.

As I will explain, understanding reasons-explanation just is to understand *practical reasoning*. Yet for many analytic philosophers, the idea that such reasoning might be anything *other* than ordinary event-causation is completely alien. Reasoning, it is commonly believed, just is a mechanism in which some content-bearing states are produced by others. Before I can fruitfully develop a non-reductive theory of *practical* reasoning and intentional action, this dogma must be exposed for what it is. In **chapter 3**, we thus temporarily depart from discussions within action theory in order to focus on the notion of *theoretical* reasoning. I will argue that even such non-practical reasoning—reasoning about what to believe—consists in a form of causality that cannot be reduced to ordinary event-causal explanation.

The reason why theoretical reasoning cannot be given a reductive treatment is that

it is an essentially self-conscious form of causality. In making an inference, we *know* that we are accepting a new belief on the basis of things we already believed before. I will argue that this link between self-knowledge and reasoning is inseparable: it is not an accidental feature of reasoning that it leads to self-knowledge. Instead, self-knowledge and reasoning are one. I begin my defense of this thesis, in the first part of chapter 3 (§§3.1-3.3), by criticizing extant accounts of self-knowledge of what we believe. These accounts attempt to accommodate the phenomenon of self-knowledge within the orthodox reductive picture to which CTA also belongs. On this view, to know that one believes p is just to be in the (justified and true) psychological state of believing 'I believe that p'. Such higher-order views thus accepts what I call the Distinct Existences assumption, or (DE): to know what one thinks is a separate act of mind from the thought itself. I will argue that accepting (DE) makes it impossible to give an adequate account of the epistemology of self-knowledge. For, I argue, self-knowledge is not limited to knowledge of what we believe, but also encompasses knowledge of why we believe it. Such *knowledge of grounds* is incompatible with (DE).

An adequate account of self-knowledge thus requires us to reject (DE) in favor of what Boyle [2011] has recently called 'reflectivism'. According to this theory, reasoning is an inherently self-conscious act. Moreover, I argue that reflectivism should be accepted because the (DE) assumption is premised on a false doctrine—the idea, famously criticized by Anscombe [1975], that the first-person pronoun 'I' is a referring expression. Anscombe's criticism of that doctrine is often unjustly swept aside. I show that the reflectivist account does justice to her negative point, while also spelling out a positive alternative conception of first-person thought. On this conception, we can drive no wedge between self-knowledge of grounds and reasoning: acquiring a new belief by reasoning *is* representing oneself as believing it.

Thus, I argue, the causality of reasoning is *not* the same as 'ordinary' event-causation. Reasoning is *spontaneous*: it rests on a kind of knowledge that is not a reflection of a state the subject is in anyway, but which instead is the subject's coming to be in that state. In acquiring such knowledge, the subject *self-determines* what she believes. In that sense, self-knowledge and reasoning are *spontaneous*.

Armed with this understanding of theoretical self-knowledge, I set out in **chapter 4** to argue that *practical* reasoning and self-knowledge are equally spontaneous, and thus irreducible to blind event-causation. I do so by giving an account of intentional action heavily inspired by Anscombe's [1963] *Intention*. Anscombe argues that intentional action is essentially characterized by practical knowledge: knowledge of what one is doing, and why one is doing it. However, in recent years some philosophers of action have tried to capture the insight that action requires practical knowledge in reductionist terms, by incorporating it within variants of CTA. I argue that this is a

mistake: a sound account of action requires that we understand practical knowledge as self-knowledge, and that is, along the lines explored in chapter 3.

Thus, I show that we must reject the practical analogue of the (DE) assumption: the idea that an agent's representation of what she is doing is an act or state *separate* from the action itself. As long as we accept (DE), the relation between what an agent thinks she is doing and what she is actually doing will be merely accidental. In §§4.2.3-4.3.1 I argue that it follows from this that CTA cannot account for the non-accidentality that is essential to practical reasoning and action-explanation, and must therefore be rejected. This is the fundamental issue which generates the cases of deviant causality we have already seen in chapter 2.

On the Anscombean account I develop, an intentional action is the agent's answer to the question of practical reasoning, i.e., 'what should I do?'. In reasoning from premises to a conclusion—an action of φ'ing—the agent represents herself *as doing* φ *because she is doing* (or wants to do) ψ. So an intentional action is 'a thought that is a movement' [Rödl 2007, p. 19]: it is a real-world happening the unity of which *is* the agent's practical knowledge. Her practical knowledge *determines* it that she is doing φ. Again, such knowledge is therefore not a reflection of a movement that the agent is performing anyway. Practical knowledge, I argue, is spontaneous and self-determining.

In **chapter 5**, I will argue that because of the spontaneous nature of intentional action, such action must be physically undetermined and free. I first argue that acting intentionally is incompatible with *universal* determinism. On the Anscombean account I defend, practical reasoning is a genuine form of causality: it explains the *occurrence* of the real-world movement which is one's intentional action. And if it is a genuine form of causality, then it must not already be determined that the agent will do φ by her 'sub-personal machinery' [McDowell 1999, p. 102]. There must be indeterminism on lower levels in order for explanations in higher-level (e.g., intentional) terms to be possible.

I defend the possibility of such higher-level causality by arguing for a *powers-based* account of causality, which has recently become increasingly popular. I show that if we think of causality as the exercise of a power or potentiality of a substance, then there can be higher-level powers which determine things not yet determined by lower-level powers. Such higher-level powers do not break the laws of nature, provided that there is lower-level indeterminism. The power to act intentionally, I suggest, is such a higher-level power.

Although the existence of a power to act implies lower-level indeterminism, I argue this does not yet imply libertarianism. For it may be that the higher-level power to act is *itself* a deterministic power—a power that can be triggered by things

as described on the higher-level itself. Existing libertarian accounts, such as Steward's [2012], fail to take note of the need to prove that the power to act cannot be a power with a trigger. This is because they appeal to the idea that action is *self-movement*, which they think implies indeterminism. I agree that action is self-movement, but show that there are different senses of self-movement, some of which do not imply higher-level indeterminism. These notions of self-movement correspond to different senses in which it is not accidental that a substance manifests a certain power. I argue that we must distinguish, among others, the self-movement of plants and animals from the rational form of self-movement exhibited by intentional action.

Finally, I provide the missing argument for the claim that the power to act intentionally is an indeterministic power. I do so by showing that the assumption that the power to act has a trigger is inconsistent with the spontaneous nature of practical knowledge. The power to act, I argue, is a self-determining power, which cannot be triggered by anything outside it. I thereby explain how our actions are neither undetermined, nor accidental, but free.

* * *

Chapter 1

Free will, luck, and action: an overview

The free will debate has traditionally been split into the opposing camps of compatibilists, hard incompatibilists, and libertarians. The position that I will defend in this thesis belongs to the last camp, for I will argue that *indeterminism* is a necessary precondition of freedom of the will, and that human agents actually possess this freedom. Libertarianism today is a troubled view, facing at least two great challenges: first, the difficulty of giving a strong argument for the incompatibility of freedom and determinism, and second, the difficulty of answering an argument often called the *luck objection*, which maintains that undetermined actions must be random happenings, outside of the agent's control. As we will see in this chapter, the two challenges are related in interesting ways. Both problems, I will argue, can be traced to the near universal adoption, among all the parties in the debate, of the so-called *causal theory of action* (CTA). That will be the main thesis of this chapter.

When we consider the landscape of the contemporary free will debate, the uniformity in the participants' understanding of the nature of *intentional action* ought to be the first thing that strikes us. Yet the existence of this shared presupposition is rarely even noticed, so obviously is it taken for granted. I will therefore provide an overview both of CTA itself, and the ways in which different accounts of free will depend on it. I will start by providing a definition of incompatibilist (libertarian) accounts of free will, and their relation to other contenders in the debate (§1.1). I will then proceed to introduce the causal theory of action, and the ways in which both compatibilist (§1.2.1) and libertarian (§1.2.2) accounts have built on it. In §1.3, I will

then examine the main objection to the possibility of incompatibilist freedom: the argument, or family of arguments, known as the luck objection.

The intuition behind the luck objection—that luck, in some sense, is opposed to freedom—is essentially sound, but should not be taken as an argument against indeterministic freedom, or so it will be argued in chapter 2. Here I will simply aim to give a reconstruction of the dialectics of the debate about the objection. As we will see, libertarians do (rightly) feel the force of the argument, and have attempted to meet it by introducing modifications to the standard libertarian picture. We will consider a number of such modified accounts (§§1.3.1-1.3.2), and see that they do not fare any better against the luck objection. If it is right that the standard libertarian picture is vulnerable to luck, then so are these modified accounts—for the modifications must make use of essentially the same resources. I will then review attempts to dispel the luck objection by appealing to the account of rational control—i.e., CTA—in common between compatibilist and libertarian accounts (§1.4)). I will show that this strategy may allow the libertarian to avoid the luck objection—but only at the cost of obscuring the motivation for the claim that indeterminism is necessary for freedom (§1.4.1).

This result may sound like good news to the compatibilist: the libertarian, it seems, cannot prove that freedom is incompatible with determinism. But the fact that the victory is so easily achieved—the libertarian cannot even give content to his demands—ought to make all participants in the debate pause: in a way, the debate is over before it started. This, I believe, is a clear sign that something has gone awry. In §1.5, I suggest that the libertarian position *can* be saved, but only by radically reconceiving it. My hypothesis, to be investigated in the following chapters, will be that CTA is an inherently compatibilist doctrine—so that contemporary libertarians attempt to defend incompatibilism after having, in effect, already bought into the opposite position.

1.1 Libertarianism and other positions in the free will debate

Let's begin by providing an overview of contemporary libertarian theories of free will. As I said, libertarianism is, roughly, the view that free will requires indeterminism, and that our actions are actually are free (and therefore undetermined). This position is opposed to several others. To begin with, there is *compatibilism*.[9] One variant of compatibilism claims that neither determinism nor indeterminism are necessary

[9] See §1.2.1 for some examples of compatibilist theories of free will.

for freedom, and so that free will exists (or may exist) either way. I will call this *neutral* compatibilism. Another position is that, while freedom is compatible with determinism, it is incompatible with indeterminism—that is, that determinism is a positive requirement for free will. This position is of special interest to us, because it is based on the so-called luck objection, an argument that will occupy us at length later in this chapter (§1.3 and further).[10] The label 'compatibilism' is, for lack of a better option, also used for those who defend this position. I will partly stick to this convention, but to prevent confusion with neutral compatibilism, I will style this position *deterministic* compatibilism.

Finally, there is the possibility of denying the reality of freedom, which comes in two varieties. According to the first position, *hard determinism*, free will does not exist because it is incompatible with determinism (to that extent, the hard determinist agrees with the libertarian), and determinism is actually true. Human actions *would* be free if they fulfilled the libertarian criteria, but this is unfortunately not the case. The second position, *hard incompatibilism*, denies that free will is possible *at all*—both under determinism and indeterminism. The hard incompatibilist thus accepts both typical libertarian arguments for the incompatibility of freedom and determinism (e.g., the Consequence Argument [van Inwagen 1986]), *and* the deterministic compatibilist's argument that indeterminism and freedom are incompatible.[11]

As it will be my ultimate aim in this thesis to defend libertarianism, it is important to get clear on the exact thesis that it affirms. Libertarians believe the future must be open: in making a choice, an agent must be able to actualize one of multiple possible courses of events. They thus believe free will is incompatible with determinism. In the contemporary debate, the notion of determinism is often taken to be equivalent to that of *universal* determinism: the view that *everything* is determined, i.e., that only one future course of events is possible. But clearly, universal determinism is not all that the libertarian ought to be worried about. For suppose that some things, e.g., the movement of some quarks located inside a distant star, are not determined. That may be the case, while it is still perfectly determined that, say, Alice will lie to her mother tomorrow at noon. So clearly, what the libertarian means is not just that free will

[10] Classic examples would seem to be Hume, who held that 'liberty, when opposed to necessity, not to constraint, is the same thing with chance' [2000 (1748)] and Hobart [1934]. There appear to be few today who actually espouse this position—most of those who now accept the luck objection end up endorsing some form of denial of freedom (see below) or suspend judgment about whether or how free will is possible (for instance, van Inwagen [2000] and [Shabo 2011]). However, Haji [2005] appears to accept both the luck objection and the view that the kind of freedom relevant to moral responsibility (which is his main interest) is at least partly compatible with determinism.

[11] Examples are Pereboom [2007] and Levy Levy [2011]. I am unsure whether the distinction between the labels 'hard determinism' and 'hard incompatibilism' is consistently used in the literature, it seems like a useful distinction to make. It distinguishes those that are conceptually in agreement with libertarians, and thus reject the luck objection, from those who do not.

is incompatible with universal determinism, but the stronger thesis that *our actions* are not determined. Universal determinism is incompatible with freedom because it *implies* that our actions are among the events in the universe which are determined. At the same time, of course, the libertarian is also not committed to the thesis of universal *in*determinism. It may be perfectly determined that some particles in the corner of the galaxy will collide a thousand years from now, without threatening our free will (perhaps there just is no way for us to prevent the collision from happening).

So it is clear that determinism, in the sense in which it is relevant to free will, is a *local* notion. The libertarian requirement is that a free action—which is an event taking place at some place and time—is undetermined. We can thus formulate the incompatibilist demand as follows:

> (**Incompatibilism**) If an agent freely φ's at time t, it was possible for her immediately before t *not* to φ at t.[12]

It is clear that this sort of possibility should be *real* possibility—it must be possible for the agent in a *physical* sense, or 'given the past and the laws of nature' as it is often put. We will return to the question of how exactly to understand such possibility in §5.2.4. For now, it is important to see that such real possibility is opposed to mere *epistemic* possibility: according to the libertarian, it is not just that we must not *know* what we are going to do in order to act freely, but rather, there must be no fact to the matter yet.

Our incompatibilist principle has obvious similarities to the famous Principle of Alternative Possibilities (PAP), formulated and attacked by Harry Frankfurt [1969]:

> (**PAP**) A person is morally responsible for what he has done only if he could have done otherwise [Frankfurt 1969, p. 829].

Indeed, since Frankfurt's formulation of the principle, many libertarians have come to espouse (PAP) as the very core of their position. But there are two important differences between (PAP) and (Incompatibilism). First, (PAP) speaks not of free will, but of moral responsibility. Although it seems true that moral responsibility requires free will, it does not follow that free will requires indeterminism *because* moral responsibility requires alternative possibilities. Either way, I prefer to stay agnostic about moral responsibility,[13] instead focusing directly on the more fundamental concept of freedom. The idea behind the demand formulated in (Incompatibilism) is simply that agents must be able to shape the future by deciding what to do. The intuitive

[12] This formula abstracts from an important issue. An action, of course, is not a momentary thing, but takes time. So to be more precise, we would have to say that what must be undetermined is the agent's *starting* to act at t. I will discuss the importance of the temporally extended character of action in §4.3.4.

[13] I thus follow the advice of Anscombe [1963, p. 11], who paraphrasing Bradley, held that 'moralism is bad for thinking'.

picture behind the requirement is that the future is like Borges's [1962] 'garden of forking paths', in which the agent decides which path to follow.

Secondly, PAP requires alternative possibilities in the sense of multiple real options for intentional action: the thesis claims that one is only morally responsible for intentionally φ'ing if one could instead have performed a different intentional action ψ. For instance, Jones cannot be responsible for voting for Smith if he could not instead have voted for Brown. That is why Frankfurt [1969] could argue that (PAP) is false. The so-called Frankfurt cases show, roughly, that it may be the case that Jones actually votes for Smith freely (or responsibly), and that it was not possible for him to vote for Brown instead, because an evil neuro-scientist would have intervened (by means of a brain implant) if Jones hadn't formed the intention to vote for Smith.[14] By contrast, (Incompatibilism) requires only that it is possible for Jones *not* to vote for Smith. This does not require Jones to be able to vote for Brown. It is enough that there is *something* that could have happened instead of the intentional action of voting for Smith: for example, Jones smashing the voting machine in a fit of anarchist rage, Jones being struck by lightning, or quite simply, his just standing in the voting booth doing nothing.

What relevance does the possibility of Jones doing nothing at t have for moral responsibility or free will? Well, if the evil neuro-scientist would also prevent all outcomes in which Jones does not press the 'Smith' button on the voting machine, then he would quite simply be forcing Jones to vote for Smith.[15] On the other hand, if any of the above mentioned alternative outcomes would have occurred, then Jones would obviously not be responsible for voting for Smith, since he didn't.[16] The enormously overcomplicated literature on this topic has taken a while to catch up to this common sense idea, and the thought that the libertarian might simply demand alternative possibilities for the future in the sense of (Incompatibilism) is now being presented as a substantial discovery under the name of 'source libertarianism' [Pereboom 2003, pp. 185-186], as opposed to the 'leeway libertarianism' of those who espouse (PAP). To me it seems that this so-called leeway incompatibilist is a straw man that did not exist before Frankfurt invented him, and that (Incompatibilism) simply expresses the natural reading of PAP's ambiguous 'could have done otherwise' clause. For it is unclear why any libertarian would ever want to claim that, e.g., someone paralyzed so as only to be able to blink with one eyelid could not freely

[14] The literature on the Frankurt-cases is enormous. See Widerker and McKenna [2003] for an overview. In the course of the debate on this topic, many libertarians have attempted to defend (PAP), which has lead to ever more complicated Frankfurt-style examples. I think the attempt to defend (PAP) is futile, but as I explain here, doing so is also unnecessary for the libertarian.

[15] In fact, it seems to me that it even goes too far to say that, in such a case, Jones performs an action of *voting* at all.

[16] See Steward [2009] for a similar view.

perform precisely that action. What the libertarian cares about is that his blinking, if it occurs, is undetermined—that he did not *have* to blink there and then. Nor does it seem that, as some surprisingly believe [McKenna 2009, §2], denying (PAP) is to depart from Borges' metaphor: by freely choosing to act, agents can still determine along which path the future unfolds.

Of course, it remains to be explained *why* exactly (Incompatibilism) is a requirement of freedom (if we rule out (PAP) as an adequate ground). One important argument for incompatibilism has been the so-called Consequence Argument, which will briefly be discussed in §1.2.1. But as we will see in §1.4.1, a variety of reasons for espousing (Incompatibilism) have recently been proposed. I will argue that none of them are convincing, and it will be my task to do better (see especially chapter 5). But to do so we must first understand why contemporary libertarian theories are not up to the task. For as I will argue, their inability to make sense of the requirement of indeterminism is due to a fundamental flaw in all their positions: that of assuming a difference between intentional action and freedom of the will. This flaw stems from the dominance of the theory of action underlying these account: the Causal Theory of Action, or CTA. I will explain this theory (§1.2), and then show how it provides the basis for modern compatibilist and libertarian accounts (§§1.2.1-1.2.2).

1.2 The Causal Theory of Action

The causal theory of action can best be understood by studying Donald Davidson's classic [1963] exposition. This seminal paper has lastingly changed the philosophy of action. Although many variants of CTA have since been developed, departing to some extent from the original, Davidson's general thesis—often summed up in the slogan that 'reasons are causes'—has become almost completely dominant, and has even been described as one of the few achievements of contemporary analytic philosophy.[17]

CTA seeks to answer the following question: what makes for the difference between an *intentional* and an *unintentional* action? Or to put the same question in Davidsonian jargon: what is the difference between a mere *bodily movement* and an *action*? If an agent raises his arm, the event that occurs seems indistinguishable from what happens when an agent's arm rises unintentionally, e.g. as the result of a queer spasm. Thus Wittgenstein's famous question arises: 'what is left over if I subtract

[17] Julia Tanney [1995, fn. 3] attributes this phrase to Millikan [1993], but interestingly, it cannot be found in that work. However, Jesse Mulder indicates that Millikan *has* described Davidson's thesis in that way in personal conversation. The exact origins of the characterization of CTA as one of modern philosophy's few achievements thus remain mysterious, though it still seems to accurately describe the mainstream opinion about the theory: the phrase is repeated in McGuire [2007].

the fact that my arm goes up from the fact that I raise my arm?'[18] [Wittgenstein 2001 [1953], §621]

Davidson's answer is that the difference lies in the *causal history* of the two events. We can summarize his views in the following way. The arm-raising, but not the arm-rising, is performed *for a reason*. The reason for which an action was performed—which Davidson calls its *primary reason*—consists of two elements: first, a *pro-attitude* towards actions of some kind, and second, a *belief* that φ (the agent's prospective action) is of that kind [Davidson 1963, p. 23]. Such a belief-desire pair[19] rationally explains (*rationalizes*) the agent's action: Alice desires to have an adventure, and believes she can achieve this by drinking what's in the bottle—so she drinks it. But the belief-desire pair is also the *cause* of the action, in the same sense as the force exerted by the cue is the cause of the ball's movement.

We can now compare the two cases of the arm-rising and the arm-raising:

1. *(Arm rising)* Alice's arm rises, caused by a queer sort of spasm.

2. *(Arm raising)* Alice desires to signal a friend and believes she can do so by raising her arm. This combination of pro-attitude and belief causes her arm to go up, while at the same time rationalizing the action.

Only in the second case is what happens (Alice's arm going up) intentional, a proper action instead of a mere bodily movement. The difference lies purely in the causal history of the two events, which are otherwise identical. But how can it be true that Alice's reason both causes and rationalizes her action? Rational justification, at first sight, seems to be completely different from causal explanation. On CTA, this difference is bridged by positing that both an agent's pro-attitude and her means-end belief are mental states, so-called *propositional attitudes*. A reason rationalizes an action in virtue of the propositional content of these states ('desiring X' and 'believing that φ is a means to X' rationalizes φ'ing). But it also causes the action because the mental state is identical to a physical state—most likely a state in the brain.[20] And

[18] As is often pointed out, it does not seem likely that Wittgenstein intended this question to have an answer. And as I will later argue (§4.3.1), Davidson's way of setting up the central problem of the philosophy of action in terms of Wittgenstein's question is mistaken—as is the seemingly innocuous assumption that there is no intrinsic difference between an arm-raising and an arm-rising.

[19] Pro-attitudes are most often thought to be desires only, although Davidson is clear that other kinds of attitudes could equally fit the bill [Davidson 1963, pp. 23-24]. The focus on belief-*desire* pairs is probably due to the popular combination of CTA with the Humean thesis that only desire can lead to action, as expressed, e.g., by M. Smith [1995, pp. 92-129]. I will follow this fashion when discussing CTA.

[20] Davidson's own view on the relation between mental and physical states is his *anomalous monism*: certain states (or rather, events) described in mental vocabulary can also be described in physical vocabulary, while these two vocabularies remain mutually irreducible [Davidson 1970]. However, anomalous monism has fallen into disregard, and many later proponents of CTA have turned to a fully reductive approach, on which mental states are type-reducible to physical states [e.g. Kim 2007].

that physical state goes on to cause, e.g., the event of the agent's arm rising.[21] So, on Davidson's view, rationalization or reasons-explanation is a 'species of ordinary causal explanation' [Davidson 1963, p. 28]. When we explain what an agent is doing, e.g., by saying 'she is doing φ *because* she desires X', the 'because' that connects the agent's desire and her action is the same as that in the explanation 'the ball moved because the cue hit it'.[22]

Although the core of CTA—that intentional actions are caused by mental states, the contents of which at the same time serve to rationalize the action—is accepted by the vast majority of contemporary philosophers, modern adherents of CTA depart in various ways from Davidson's account in *Actions, reasons and causes*. For instance, Searle argues that the causation of action by reason states must be a continuous matter: when Alice is raising her arm intentionally, the relevant mental state (an 'intention-in-action') must continue to exert causal influence on the movement of her arm during that period [Searle 2003, pp. 46-49].[23] Again, others [e.g. Bratman 1987] argue that the kind of mental states required for intentionality are different than those Davidson originally proposed: not belief-desire pairs, but rather *intentions*.

In fact, Davidson himself later admitted the need for adopting intentions as mental states that play a crucial role in the production of action. Intentions play a mediating role between an agent's reasons (her belief-desire pairs, of which an agent might have any number) and her actions:

> It is a reason for action that the action is believed to have some desirable characteristic, but the fact that the action is performed represents a further judgement that the desirable characteristic was enough to act on—that other considerations did not outweigh it. [Davidson 1978, p. 98]

This 'further judgement' is the intention. Consider an example. Suppose that Alice, being hungry, has reasons for two distinct actions: she desires to eat something sweet, believing she can do so by eating this delicious cake, and she desires to lose weight, believing she can do so by eating only low-fat food (say, the crackers in the cupboard). All things considered, she desires most to lose weight, and forms a corresponding intention. This is rationally understandable, given her that she wanted to have the crackers more than the cake. At the same time, Alice's intention is causally explained by the fact that her desire to lose weight was stronger than her desire for the cake.

But no matter which states are precisely posited, all these variants of CTA share a common picture which is far more important than the differences between them.

[21] As Davidson [1970, p. 30] points out, when we speak of a mental state as a cause, this should be understood as the event of the *onslaught* (i.e., onset) of the state causing some other event (e.g., the onset of a different mental state, or an action).

[22] I will explain more concerning Davidson's motivation for this view in §2.1.2 and §2.2.1.

[23] Compare Frankfurt [1978], who criticizes Davidson for the temporal separation of cause and effect.

On all variants, an action or intention is the rational *and* causal consequence of which way the scale of reasons tips. Reason-states are physical forces, each almost literally pushing the agent in a different direction. As Davidson says: 'if reasons are causes, it is natural to suppose that the strongest reasons are the strongest causes' [Davidson 2001a, p. xvi]. So if the net force of all reason-states favors φ'ing, this causes the agent to φ (or to form an intention to φ). McDowell [1981, p. 213] dubs this a 'quasi-hydraulic conception of how reason explanations account for action'.[24]

Given a certain balance of reasons, it thus seems *necessary* that the agent will φ (or at least that she will form an intention to do so[25]). However, a caveat should be made here. It is possible that Alice will act weak-willed or *akratically*, i.e., that by some irrational causal disturbance, she will *not* act in accordance with what she judges to be her best reasons [e.g. Davidson 1969, Kalis 2011].[26] So even if the balance of reasons favors φ'ing, it may still not follow that the agent will φ. However, this is only because when we only consider her reasons, we do not take into account all the causes that exert force on the agent.

1.2.1 CTA and compatibilism

Compatibilists, of course, reject (Incompatibilism). That is, they believe an action can be free even if it was causally determined that the agent would perform it. It may be that there are certain other conditions that must be fulfilled in order for the action to count as free. For instance, classical compatibilists argued that—since freedom has nothing to do with whether the action is determined or not—the true contrast between free and unfree action must lie in whether the agent is coerced or not. On that picture, to act freely is to be unhindered in the pursuit of what one desires.[27] Other compatibilists argue that an action must not (or not only) be un-coerced in

[24] He then goes on to dismiss it as 'a radical misconception'. Smith [1995, pp. 101-102] elaborates on the meaning of McDowell's metaphor: '... the mind is an arena where various forces ... get channeled in certain directions and ultimately combine together to produce a resultant force ...'. Smith denies that he is in fact committed to such a picture, but his 'dispositional' account of reasons-explanation seems to be merely a causal theory in disguise.

[25] Davidson believes it is possible for an agent to intend to have so-called 'pure intentions': an intention to φ which never goes on to cause an action of φ'ing [Davidson 1978]. If that is right, the question is raised: what determines whether an agent will act on an intention or not? Given the causal picture of action, it seems plausible to say that intentions, just as desires, also come in different strengths: if an agent does not act on her intention to φ, some *other* force must have pulled her in a different direction.

[26] The possibility of akrasia is a notoriously difficult problem for Davidson. Weakness of will can only occur if an a-rational causal force disturbs the balance of reasons. But paradoxically, if that force is a-rational, it seems that the agent does not act for a reason, and hence not intentionally. Yet for Davidson, akratic action is intentionally acting against your best judgment. Davidson only resolves this paradox by introducing different 'partitions' of the mind, which causally interact in the way that the beliefs and desires of distinct agents can also interact. He argues the resulting behavior is then still intentional [Davidson 1982, pp. 180-181].

[27] A paradigmatic example of this is Hobbes, who argued that an agent's freedom consists in 'that he finds no stop in doing what he has the will, desire, or inclination to do' [Hobbes 1651/2012, p. 117].

order to be free: what is necessary is that the action is the result of the agent's *own* desires, in some sense.[28] For still others, nothing additional is required: when one acts on the basis of one's reasons, that already is to act freely in some sense (although, of course, one might then still be unfree in the sense of being coerced). That seems to be Davidson's own view.[29]

The varieties of compatibilism are thus many, and well-known. It will not be my purpose to give an overview of all these positions.[30] Rather, I want to draw attention to the importance of CTA[31] to many contemporary compatibilist theories. The best way to do so is to consider the compatibilist's response to what has arguably been the most important incompatibilist argument in the modern debate: Van Inwagen's Consequence Argument. Here is an intuitive formulation of the argument:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. [van Inwagen 1986, p. 16]

As Bishop notes, we should presumably understand an act's not being up to us here as its not being 'an exercise of the agent's control' [Bishop 1989, p. 53].[32] That becomes clear when we consider that Van Inwagen's more formal presentation of the argument depends on the modal operator N, which he asks us to read as 'no one has or had a choice about . . . ' [van Inwagen 1986, p. 93]. The argument then proceeds to show that, since no one ever had a choice about the laws of nature and the initial conditions of the universe s, and (under determinism) s entails that an agent would do p, it follows that Np—i.e., 'no one ever had a choice about doing p'.[33] Thus, the Consequence Argument aims to show that, under (universal) determinism, our present behavior would be something about which we do not have a choice.

Many find this an intuitively convincing argument. And it has sparked much

[28] For Frankfurt [1971], e.g., the desires must be endorsed by being the object of a second-order desire of the agent's.

[29] Compare Davidson [1970, 1973]. See §1.5 for more discussion of Davidson's position.

[30] For an extensive overview, see McKenna [2009].

[31] Or its primitive precursors, as for example in Hobbes. Heath [2008, pp. 11-15] illustrates the similarities between Hobbes' account and contemporary theories of action.

[32] Bishop equates an act that is 'up to us' with an action's being under the agent's control, and he equates the latter with its being an intentional action. He thus interprets the Consequence Argument as an argument for the thesis that intentional action would be impossible under determinism—and I believe rightly so. This issue is usually ignored in discussions about Van Inwagen's argument: it is often supposed that its aim is to show that our intentional actions would not be 'up to us' under determinism. It is never explicated what the relevant sense of 'up to us' could be, if it is not to be equated with the control that an agent has over her intentional actions. In §1.5, I argue that this implicit separation of intentionality and an action's being 'up to us' is responsible for much confusion in the contemporary debate. Steward [2012, p. 31] also remarks on this unfortunate separation of the idea of action that is up to us from that of action *simpliciter* in discussions of the Consequence Argument.

[33] The formal details of the argument need not bother us here. See Steward [2012a, pp. 26-32] for an overview of the debate surrounding the formal issues.

debate, with compatibilists trying to defuse the inference rules on which the formal renderings of the argument depend. However, we should note that the argument depends on certain controversial assumptions about the nature of *choice*, and hence, of *intentional action*.[34] Van Inwagen just assumes that we cannot have a choice about something that is causally determined.[35] That assumption becomes uncompelling when we are presented with an *account* of choice which shows that choice *is* compatible with necessitation by prior states. And that is precisely what CTA does present us with. As Bishop says:

> . . . the Consequence Argument is false because the core claim of the Causal Theory of Action is true. [. . .] if the agent's behavior, which is the *deductive* consequence of these unavoidable states of affairs, is also the *causal* consequence of the right kind of states of the agent, the agent's behavior will fulfill conditions sufficient for it to count as action. If facts about the remote past cause present behavior via the right sort of causal chain, the causal consequence of what is unavoidable will actually constitute a case of something that comes about through agent-control. [Bishop 1989, p. 57]

According to Bishop, this even shows that *any* compatibilist account must embrace CTA [Bishop 1989, p. 59]. That may be a slight exaggeration: there have been some philosophers who defend compatibilism without accepting CTA.[36] But I agree with Bishop that if CTA is true, the Consequence Argument loses its bite.[37] That, it seems, is why CTA (in its many varieties) has become the foundation of almost all modern formulations of compatibilism.

1.2.2 Event-causal libertarianism

Let us now investigate how contemporary libertarian accounts make use of the foundations that CTA provides. The simplest, and arguably most popular account is known in the literature as *event-causal* libertarianism.[38] Given CTA's view of what intentional action is, the event-causal libertarian aims to answer the question what

[34] I am assuming, I think uncontroversially, that intentional action is action that the agent has chosen to do. At least for Davidson [1978], an intention or action is the conclusion of practical reasoning, and so the result of an agent's considering the question what to do—her making a choice.

[35] This becomes explicit in one of his inference rules, α, which states that $\Box p \vdash \mathrm{N}p$.

[36] One famous example is Von Wright [1974]. I argue against non-causal positions such as his in §5.1.1. Moreover, Markosian [1999] has argued for a compatibilism on the basis of agent causation. However, I think that his compatibilist version of agent causation has the same problems of libertarian agent-causalists: it cannot explain how the resulting action is connected to the agent's reasons except by implicitly still relying on CTA. See §1.3.2.

[37] As Bishop [1989, p. 58] notes, it would be a weak reply to insist that CTA cannot be true because the Consequence Argument is correct—at the very least, that would require spelling out in what way CTA fails as an account of choice, or some defense of the assumption that choice is incompatible with determinism. That is the challenge I will take up in the coming chapters.

[38] Prominent event-causal libertarians include Kane [1996], van Inwagen [1986], Balaguer [2010]. Recently, Franklin [2011a] has provided an especially clear formulation of the position.

a *free* action is. Although CTA is also the preferred account of compatibilists, it can be made to accommodate the incompatibilist's demands with some small changes. Assuming that a theory of free will must conform to (Incompatibilism), how does the event-causal libertarian go about 'transforming a mere action into a free action' [Franklin 2011a, p. 203]?

Recall our example of Alice, who was deliberating whether to have cake or crackers. Since her desire to lose weight was stronger than her desire to eat something sweet, she ate the crackers. And given the relative strength of her reasons, it was necessary that she would do that. There was only one action possible for Alice.[39] (Incompatibilism) is thus false, and so according to the libertarian, Alice was not free.

Now the event-causal libertarian proposes the following: reason-states are *indeterministic* causes of actions.[40] That is, the relative strength of each reason-state corresponds to an *objective probability* that it will cause an action. And since it was natural to suppose that the strongest reasons are the strongest causes, it is now plausible to conceive of stronger reasons as having a higher objective probability. If Alice's desire to lose weight is twice as strong as her desire to eat something sweet, she is twice as *likely* to have the crackers. The probabilities of each reason collectively add up to 1.[41] So instead of the deterministic 'winner takes all' scenario, we have a picture on which even a less strong desire can end up as the winner. When Alice is caused to act by one of her reason-states, 'it did not *have* to cause [her action]; it just *did*' [van Inwagen 1986, p. 141]. Before she acted, any other reason (or combination of reasons) *might* have produced an action. Thus, it is easy to see, (Incompatibilism) is satisfied: Alice's actions are not merely intentional, but, so the event-causal libertarian claims, also free.

As mentioned, event-causalism is the most popular view among contemporary libertarians. Although their views differ in respect to exactly which mental states cause which other states or actions, the essential common ingredient is to require that, instead of the strongest reasons being causally sufficient, reasons have objective probabilities corresponding to their relative strength. We should be very clear on

[39] The caveat introduced above (§1.2)—the possibility that the agent will act akratically—does not do anything to alleviate the incompatibilist's worries here. For even if we cannot *know*, given only the balance of the agent's reasons, that she will φ, it *is* already settled whether or not she will do that. The caveat only shows that making a (deterministic) prediction of an agent's action sometimes requires knowledge of not just the balance of reasons, but also of some other causal factors.

[40] Anscombe [1971] convincingly argues that necessitation or sufficiency is, contrary to what many philosophers still think, not part of the concept of causation. Hence, indeterministic insufficient causes are conceptually possible at least—and modern physics seems to show they are actual, too, quite apart from the question whether quantum mechanics has anything to do with free will. See §5.2.1 for more.

[41] If the event-causal libertarian, like Davidson, wants to take seriously the possibility that an agent may fail to act at all (for instance, because she forms a pure intention), she might say that these probabilities are relative to the assumption that she *will* act intentionally: *if* Alice acts, it is twice as likely that she will φ than that she will ψ.

what this move amounts to. It is *not* a modification of CTA—i.e., it is not that the libertarian is saying that Davidson's view on intentional action is mistaken. Rather, the libertarian accepts CTA as an account of intentional action, and then insists that given that account, it is conceptually possible for an intentional action to come about indeterministically. And if that is actually the case for any of our actions, it will be not just intentional, but also free. The reason for the popularity of this view, no doubt, is exactly that it requires so little in addition to CTA. If rationalization is a species of causal explanation, and if (as modern physics seems to show, see fn. 40) causation can be indeterministic, then actions can, *prima facie*, have indeterminstic causes. At most, the libertarian's opponents can argue that indeterministic causation is not, as a matter of empirical fact, present in the production of our actions.

1.3 The luck objection

The event-causal libertarian recipe for 'transforming a mere action into a free action' [Franklin 2011a, p. 203] makes use of the smallest possible additional requirement to CTA: that of indeterministic causal connections between reason states and the resulting actions. Small as this additional requirement might be, many contemporary philosophers argue that it is enough to render actions caused in accordance with this recipe *un*free. The argument to this effect has recently received much attention and can be formulated in a variety of ways. We will refer to it as the argument from luck or the *luck objection*. This objection is of great interest to us, because it will reveal the extent to which the assumption of CTA shapes and limits the debate. I will thus use it as a sort of case study in order to highlight what is, according to us, mistaken about the contemporary debate about free will.

In order to understand the general worry behind the luck objection, consider Mele's metaphorical representation of the event-causal picture:

> As soon as any agent . . . judges it best to A, objective probabilities for the various decisions open to the agent are set, and the probability of a decision to A is very high. Larger probabilities get a correspondingly larger segment of a tiny indeterministic neural roulette wheel in the agent's head than do smaller probabilities. A tiny neural ball bounces along the wheel; its landing in a particular segment is the agent's making the corresponding decision. [Mele 2006, p. 8]

What this analogy is supposed to make clear is that what an agent ends up doing, for the event-causal libertarian, is in the most literal sense a matter of *luck*. And intuitively, what is a matter of luck or chance is not free. For example, the outcome of a die-roll (supposing it is an objectively indeterministic process) is a matter of luck. Say an agent throws a six. Intuitively, this was not *up to her*—it was not under her

control. In the same sense, many critics of libertarianism want to argue, it is not up to the agent which action she will perform, given a certain balance of reasons and the accompanying probabilities. So the result of the neural roulette game cannot be free.

Many different thought experiments expressing this general worry abound in the literature. However, it seems fair to say that each of them attempts to establish the conclusion that undetermined action cannot be free by arguing for (or assuming) the following two claims:[42]

1. If an action is undetermined, it is a matter of luck.
2. If an action is a matter of luck, it is unfree.

This, then, is the most general formulation of the luck objection. I will not aim here to give a comprehensive overview of all the instances of this argument form, as little that we have to say about the libertarian's ability to cope with the objection depends on any one way of phrasing it. However, we will now consider a paradigmatic instance of the argument.[43]

A recent precise formulation of the argument requires a so-called *contrastive explanation* of the agent's decision.[44] The thought is that, if an agent's decision is undetermined, then there is no accounting for the fact that she decides to do one thing *instead of* doing another:

1. If an agent's φ'ing is undetermined, it was equally possible for her to ψ instead.[45]
2. If it was equally possible for her to ψ instead, nothing explains why she φ's-instead-of-ψ's.
3. If nothing explains why she φ's-instead-of-ψ's, it is a matter of luck that she φ's.
4. If an agent's φ'ing is a matter of luck, it is not free.
5. Hence, indeterministically caused actions are not free.[46]

[42] Franklin [2012, p. 410] disagrees. He claims that the luck objection is to be distinguished from, for example, Shabo's [2011] 'Assimilation Argument. *Pace* Franklin, to me it seems obvious that Shabo's claim that undetermined actions are indistinguishable from random events is yet another way of describing premise (1) as we give it here.

[43] The Replay Argument, one of the most popular versions of the luck objection at the moment, will be considered in §1.4.

[44] The contrastive formulation of the luck objection is endorsed by, e.g., Ginet [1997], Haji [2005], Mele [2006], Levy [2011].

[45] ψ should here be understood as including the possibility of not doing anything intentionally at all, for reasons explained in §1.1.

[46] This way of making the contrastive formulation explicit is adapted from Franklin's [2011, p. 220] rendering of Mele's [2006] argument.

On a deterministic picture, if an agent's reasons favor φ'ing, this makes it inevitable that the agent will decide to φ (assuming that she will not, e.g., be struck by lightning). The balance of her reasons thus explains fully why she φ's-instead-of-ψ's. On the event-causal libertarian account, however, once the relative force of each reason is set, 'it remains open whether [a particular] decision will occur, and whether it will is not settled by anything about the agent—whether it be states or events in which the agent is involved, or the agent herself' [Pereboom 2007, p. 102]. In that sense, then, nothing about the agent settles which decision she will make, and so she is not in control of whichever decision she does end up making. It seems that one's reason for acting, on the indeterministic picture, 'cannot explain what it is supposed to explain, namely *why I did what I did rather than an alternative that was causally open to me*. It says I did it for certain reasons, but does not explain why I didn't decide not to do it for other reasons' [Nagel 1986, p. 116]. And '[w]ithout such an explanation . . . the choice [the agent] makes is not strongly connected to her prior states including her desires, beliefs, values, and so forth' [Haji 2000, p. 223].

In defending themselves against the luck objection, libertarians of course have two options: to attack the premise that establishes that the agent's action is a matter of luck, or the premise that concludes from this that it cannot be free. In the contemporary debate, 'luck' is simply taken to be synonymous with indeterminism, and libertarian defensive attempts therefore focus on the second premise. The task for the libertarian, then, is often conceived as coming up with additional conditions for freedom that would serve to undermine the premise that anything undetermined cannot be under an agent's control (though see §1.4 for a different, and more promising reply). I will now investigate two attempts to come up with such additional requirements: Kane's elaborate account, which stays purely within the event-causal framework, and the agent-causal proposals of Clarke and O'Connor, that seek to rescue incompatibilist freedom from luck by supplementing event-causalism with a different kind of causation.

1.3.1 Kane's event-causal reply

Robert Kane has composed a complex variation on the event-causal libertarian theme, largely in order to rebut the luck objection.[47] Kane asks us to imagine a man who 'makes an effort' to break a table by striking it. Even if it is undetermined whether he will succeed (we are supposedly imagining some kind of indeterministic glass), *if* the table breaks, it would be a 'lame excuse' to insist that 'chance did it, not me'

[47] Kane's position has developed over the years, as expressed in a number of publications [e.g. Kane 1996, 1985, 1999]. All references to Kane's position in this thesis will be made to his writings in Fischer, Kane, *et al.* [2007], which offers a concise overview of his latest views.

[Kane 2007a, p. 27]. Kane subsequently models his account around this example. He proposes that when agents must choose between φ'ing or ψ'ing, they simultaneously engage in *two* efforts: an effort to decide to φ (based on certain desires), and an effort to decide to ψ (based on different desires). The strength of each effort is undetermined; only when the efforts are over and the agent decides to either φ or ψ is it settled which is the strongest. We are to imagine that both efforts are simultaneously undertaken in an exercise of 'parallel processing'. Supposing the agent actually decides to φ, it was still *possible* that her effort to ψ would have come out on top. But as in the table-breaking case, it would be a lame excuse to say that 'chance decided to φ', for it was the agent's *effort* that caused the decision. [Kane 2007b, p. 173]

This is a puzzling response to the luck objection, for it is completely unclear what purpose is served by replacing the struggle between multiple reason states (as on the simplest variety of event-causal libertarianism) by a struggle between two 'efforts to decide'. Kane says the following, as a direct response to the luck objection posed by Pereboom [2007, p. 102]:

> . . . in making efforts of will to choose in terms of their reasons and motives, agents do play a causal role in bringing about their choices over and above the causal role played by their reasons, motives, intentions, and other mental states alone. In other words, it's not as if the agents sit back and watch while the reasons or motives cause the choice. It is rather that, by making efforts, the agents actively bring about the choice for the reasons and motives. Efforts are different from desires and other motives in this respect [. . .] because efforts are actions of the agents and not merely states. [Kane 2007b, p. 174]

'Efforts', according to Kane, are different from, e.g., desires, in that they are 'active'—they are 'actions' and not merely states. Yet in what sense can Kane's efforts be said to be 'actions'? The answer is that they are mental events that are caused and rationalized by previous reason states, such as desires. But in that sense, they are perfectly analogous to desires themselves on the unmodified event-causal account. And Kane accepts that on the unmodified account, the result of the causal struggle between desires is just a matter of luck, and therefore not free. So why should the efforts be 'active' in any stronger sense than the resultant intentional action is on the unmodified account? The answer cannot be, e.g., that 'efforts' are *events* instead of states, because an agent's coming to have (or having) a certain desire is, on the Davidsonian picture, equally an event. Nor can the difference be that Kane's efforts are actions in the sense of being *rational* or *intentional* events: if they are, then they are so precisely in virtue of being caused by previous reason states or events that engage in a causal struggle.

The contrast between reason states and 'efforts' is thus quite empty. In effect,

Kane's efforts seem to be nothing more than the reification of the causal struggle between desires on the original event-causal picture: where the standard event-causalist says that decisions or intentions are caused by the indeterministic causal struggle between desires, Kane views that causal struggle as *itself* an event—which must, of course, have its own causal connection to preceding and subsequent events (the agent's coming to desire something, and the agent's coming to decide what to do). So Kane's attempt to avoid the luck objection by introducing simultaneous 'efforts to decide' changes nothing to the original picture in which we have simultaneously existing reasons engaged in a causal struggle. If the latter does not serve to bring the end result of the interplay between forces 'under the agent's control', as the luck objection claims, then surely, neither does the former.

The lesson to learn is that it does not seem to matter what we *call* the states or events that occupy the places between the indeterministic causal connections—'efforts', 'desires', or whatever else suits the libertarian's fancy. For they will be nothing more than mental states or mental events with a certain propositional content, which probabilistically cause other mental events. If the relation of probabilistic causation of mental event X by mental event Y does not suffice to render event X 'active' or 'up to the agent', it does not matter how many such connections we have. The luck objection appears to shows that an action cannot be free if it is just a random result of prior states of the agent's. If one accepts that argument, the problem cannot be fixed by simply adding yet more links of the same kind to the causal chain.

1.3.2 Agent-causal replies

But perhaps the luck objection can be overcome by adding a *different* kind of link to the causal chain that the libertarian envisions as leading to free action. This is the aim of so-called *agent-causal* accounts of free will.[48] The two most prominent agent-causal accounts in the literature are those of Clarke [1996, 2003, 2005] and O'Connor [2002, 2005].[49] We will review these in turn, and argue that both accounts fail as a response to the Luck Objection for the same reasons—and for similar reasons as Kane's proposal.

Clarke's theory departs from the ordinary event-causal libertarian picture dis-

[48] It should be noted that Clarke, one of the agent-causal theorists is discussed in this section, at least does not develop his account in order to rebut the luck objection. As we will see, he does not believe that the luck objection is cause for concern for even the ordinary event-causal libertarian, for reasons we shall examine in §1.4. Rather, Clarke thinks agent causation is an answer to the 'problem of enhanced control' (§1.4.1). And it is unclear whether O'Connor espouses his version of agent causation in order to overcome the luck objection, the problem of enhanced control, or both. However, there are certainly others (e.g., Pereboom [2007], Levy [2011]) who treat the agent-causal proposal as an (in the end) unsuccessful attempt to solve the problem of luck.

[49] Agent-causal theories have a modern precursor in, for example, Chisholm [1976].

cussed so far. He regards indeterministic causation by reason-states as a necessary, but by itself insufficient condition for free action. What is further required is that 'an agent is in a strict and literal sense an originator, an initiator, an ultimate source of her directly free action' [Clarke 2003, p. 134]. This condition is satisfied by hypothesizing that agents have an agent-causal ability to produce actions *themselves*: a power that 'does not consist in causation by events' [Clarke 2003, p. 134] such as the onset of certain reasons, but which they possess 'fundamentally as a substance' [Pereboom 2002, p. 40]. In short, the picture is the following. When an agent acts freely, his action is indeterministically caused in the event-causal way. This must secure that the produced event is an intentional *action*, instead of a non-purposive happening. At the same time, and in addition to this event causation, the agent *also* exercises his agent-causal power. This must secure that the agent is the 'ultimate source of her directly free action'. Clarke thus views exercises of this power as exercises of 'direct active control'. However, this does not mean that the action is causally overdetermined. It is rather that, 'as a matter of nomological necessity' [Clarke 2003, p. 136], exercises of the agent-causal power always occur concurrently with the indeterministic event causation by reasons. Just as '[a]n increase in temperature and an influx of oxygen may both be needed . . . to cause combustion' [Clarke 2003, p. 136], so agent causation and event causation must occur together to produce an action—the two cannot come apart as a matter of natural law.[50]

For O'Connor, as for Clarke, agents possess a power that is fundamentally different from event causation, the exercise of which is uncaused by prior events, and which is an exercise of self-determination. This agent-causal power produces an 'action triggering intention', which then goes on to causally produce a further 'concurrent intention' which exists throughout the action and causally sustains it. Confusingly, O'Connor's theory of action is sometimes dubbed 'non-causal' [O'Connor 2005, Haji 2005, p. 331]. It is non-causal in the sense that the 'action triggering intention' is not event-caused (though it is agent-caused). However, it seems clear that his account *does* depend on the causal theory of action, since the concurrent intention's causing and rationalizing the action-event is that which renders it intentional. In that sense, O'Connor is firmly in Davidsonian territory. However, on CTA, the intention itself inherits its rationalizing capacity from the reasons that previously caused and rationalized *it*. And since on O'Connors account, the intention is *not* caused by the agent's prior reasons, but agent-caused instead, he must accommodate the rationalization-aspect of action explanation differently. He does this by including the agent's reason for action in the content of the action triggering intention, which is of the form 'to *A*

[50] It must be noted that Clarke doubts whether agent causation actually exists. See Clarke [2003, pp. 196-212].

for reason *R*1' [O'Connor 2005, p. 351]. Prior to agent-causing an intention, the agent was 'aware of reason *R*1 while deliberating' [O'Connor 2005, p. 351], and this fact explains why the agent formed the corresponding intention. Which reasons can enter into the content of the intention thus depends on which reasons the agent was aware of beforehand: 'coming to recognize a reason to act induces or elevates an objective propensity for me to initiate the behavior' [O'Connor 2005, p. 353]. agent causation is thus allegedly a capacity 'probabilistically *structured*' by reasons [O'Connor 2005, p. 353].

Many philosophers respond to the agent-causal proposal by offering 'the incredulous stare': they regard causation by substances, instead of 'ordinary' event causation, as a mysterious, quasi-dualist affair, or at least as a speculative hypothesis that, as far as we can see, has no prospects of vindication by physics. In reply to this, Clarke [2003, pp. 185-218] and Lowe [2008, pp. 143-147] have suggested that agent causation—or more precisely, substance causation—might actually be a ubiquitous phenomenon.[51] Perhaps substances, animate and inanimate, possess *powers* or *dispositions*, the exercise of which is the substance's causing a particular effect. Then we can 'translate' statements about event causation (e.g., 'the explosion of the bomb caused the collapse of the bridge' [Lowe 2008, p. 145]) into statements about substance causation ('the bomb caused the collapse of the bridge': it is correct to assert the former, 'because (we suppose) it was the bomb that, *by exploding* [emphasis D.O.], caused the collapse of the bridge' [Lowe 2008, p. 145]. 'Exploding' is a power that bombs possess, and this particular event of explosion was caused by the bomb by exercising that power. Of course, every statement about substance causation is equally translatable into a statement about event causation in the same way. But Lowe argues that substance causation is the fundamental kind of causation. I will not go into questions about the metaphysics of causation here (although see 5.2), but if Lowe is right, this would straightforwardly rebut the charge of mysteriousness.

It would, however, make it completely mysterious how the introduction of agent causation is supposed to help the libertarian to overcome the luck objection: if the indeterministic radioactive decay of a radium atom can be substance-caused, then saying that human actions are substance-caused does nothing, by itself, to alleviate our worry that such actions would be random, inexplicable flukes. If anything, it would increase that worry, for indeterministic processes like the decay of radium are paradigm examples of uncontrolled random events. So substance causation conceived in this way does nothing to show that 'the choice [the agent] makes' '. . . is strongly connected to her prior states including her desires, beliefs, values, and so

[51] As we will see in §5.2, I am sympathetic to this thought. However, as we will now see, the challenge is to give a coherent account of how such substance-causal powers can be *rational*, i.e., what it is for them to be exercised *for a reason*.

forth' [Haji 2000, p. 223], or that it is something about the agent herself that settles which choice she makes.

As Clarke [2003, pp. 193-196] notes, if substance causation itself is metaphysically innocent and possibly ubiquitous, agent causation (substance causation by human agents) must have specific additional characteristics that renders its exercise free. These characteristics, of course, must be that agent causation is an indeterministic and, unlike radium's power to decay, *rational* power.[52] Pereboom agrees: the agent-causal libertarian could solve the luck objection if it were true 'that the agent-cause's role is not independent of [her] reasons . . . but rather that an aspect of her agent-causal power is the capacity to consider and weigh reasons, and thereby to guide her causing of choices' [Pereboom 2002, p. 67]

The problem is that the agent-causal capacity that agent-causalists imagine does not appear to be a rational power at all—or so I will now argue.[53] On O'Connor's view, the agent causes an intention, the content of which contains a reason. Now, that the resulting state mentions a reason does not make the power a rational one: if reasons are mental states identical to or supervenient on physical states, it is easy to imagine that it could be induced by a non-rational power (i.e., the influx of electric current in the brain, or some such). In other words, that the agent-causal power results in a state that may go on to rationalize actions, as per ordinary CTA, does not mean that the agent exercises that power *itself* for a reason. O'Connor claims that the exercise of agent-causality *is* explained by the agent's reason because, prior to the agent's causing the intention, the agent deliberated and was aware of the same reason that is part of the resulting intention's content. This explains the action because awareness of a reason 'elevates an objective propensity . . . to initiate the behavior' [O'Connor 2005, p. 353]. Here O'Connor is vulnerable to a difficult question: how is it that awareness of a reason increases the objective probability that an agent will perform a certain action? For event-causal libertarianism, this is because the strength of a reason corresponds to its likelihood of becoming causally active. But on O'Connor's account, that explanation is unavailable.

And on Clarke's view, too, the agent's exercise of her power does not itself appear to be performed for a reason. Acting for a reason, on Clarke's own admission, is to be analyzed in event-causal terms. The exercise of agent-causality is *concurrent*

[52] Compare part of Lowe's [2008, p. 194] response to the luck objection: the substance-causal power of choice is 'always informed by or responsive to reasons for action. In this respect, it is utterly unlike any mere chance event, such as the fall of a die or the decay of a radium atom'. Like Clarke and O'Connor, Lowe claims that his conception of an agent's power of choice ensures that it is a rational power, but this can be doubted on the basis of the same arguments advanced below. We return to this issue of what it would be for a substance-causal power to be intrinsically rational in chapter 5, especially in §5.4

[53] Levy [2011, pp. 63-76] argues similarly, though he does not focus on the role that the 'objective propensities' play for O'Connor, nor on the problematic ubiquity of agent causation illustrated above.

with, and not itself part of, the event-causal relation between reason-states and action. Perhaps it is true in virtue of the event-causal relation between reasons and action that the agent acted for a reason, *and* true in virtue of the agent's exercise of her agent-causal power that 'the agent' performed it—but as explained above, that the agent performed it does not serve to render the action free, since even inanimate substances may be indeterministic substance causes.

We are thus forced to conclude that the agent-causal libertarian fails to establish agent causation as a rational power that provides for a strong tie between the agent's reasons and her choice: '[t]he agent's reasons bring her to the point at which the agent-causal power is exercised, but they are unable to explain the final agent-causal push *itself*' [Levy 2011, p. 67]. In that sense, what the agent ends up doing seems to remain a matter of luck that is beyond the influence of her faculty for rational choice.[54]

1.4 A reply and a dilemma for the libertarian

In the previous section, we have seen that modifications of the standard event-causal account do not fare any better against the luck objection than the original. The resources that the event-causalists has are limited to inserting new links of the same kind (e.g., Kane's efforts of will) in the causal chain, for which the same problem then arises: is it not just a matter of luck, outside of the agent's influence, that the causal chain unfolds in this way—e.g., is it not a matter of luck that the agent's effort to ϕ turns out to be stronger than her effort to ψ? The agent-causalist aims to broaden his resources by including a new type of link in the causal chain—the agent-causal push. But he is unable to explain what is relevantly different about this kind of link: *how* the agent-causal power is exercised seems to be a matter of luck in the same way as the result of the causal struggle between mental states for the event-causalist.

Interestingly, then, both modifications of standard event-causal libertarianism espouse the same model of what it is for an agent to exert rational control over an event (an action or decision): that is for the event to be caused in a certain kind of way by preexisting entities of a certain type (i.e., the agent's mental states). So when it is argued, by means of the luck objection, that the outcome of the causal struggle between desires is not under the agent's control, their only possible solution is to insist on adding links in the causal chain before these desires—links that are

[54] That is not to say that the idea of a rational power to act cannot in principle solve the luck objection—in chapter 5 I will argue that it can. But then we must understand what it is for an agent to exercise a power for a reason, and not as a matter of luck. I have tried to show that contemporary agent-causalists remain unable to explain this because they conceive of what it is for something to happen for a reason, and of what it is for an agent to deliberate, along the lines of CTA.

supposed to confer 'control', 'activity' or freedom on those desires in the same way as they themselves were supposed to confer it on the resulting action. If we suppose that indeterministic causation of an action by desires or intentions does not confer freedom or control, it can come as no surprise that additional causal factors inserted before the desire or intention equally fail to do so.

On the other hand, since all participants in the debate—libertarians and their critics alike—*do* espouse this model of rational control provided by CTA, we should ask what reason there is to accept the accusation of the luck objection in the first place, assuming that picture. Since CTA is precisely an *account* of rational control, it seems that any theory that conforms to it (like the event-causal libertarian theory) cannot be faulted for *lacking* such control. As Clarke says:[55]:

> An event-causal theory of action offers a plausible account of [direct active control], one that allows that we may have genuine action in both deterministic and indeterministic worlds. The thesis that an event that is an action is so in virtue of being nondeviantly[56] caused by certain agent-involving events strongly suggests, if not outright implies, that the action's being so caused is what *constitutes* [emphasis D.O.] the agent's exercising direct active control over that event [. . .]. Direct active control is exercised, then, when one acts, not when one undergoes earlier changes. [Clarke 2003, p. 76]

So when Pereboom [2007, p. 102] complains that once the 'relevant causal conditions antecedent to a decision' are set, 'the agent has no further causal role' in settling whether a particular decision will be made, the answer is simple: the agent settles what will happen *when* she makes a decision, precisely *by* making that decision. The causation of the decision by the agent's reason states *is* the agent's settling what she will do. There simply is no possible other role for the agent to play that can be missing due to indeterminism. This reaction to the luck objection thus mirrors Bishop's reply to the Consequence Argument on behalf of the compatibilist. Both the Consequence Argument and the luck objection try to show that an event cannot be under an agent's rational control if it is determined or undetermined, respectively. And in both cases the answer is to give an account of such control—CTA—which undermines the intuitive assumptions underlying the argument. So it seems that, if the compatibilist is allowed to use CTA to rebut the Consequence Argument, the event-causal libertarian can do the same to rebut the luck objection.

However, some proponents of the luck objection have attempted to show that this simple reply does not succeed. For example, according to Shabo:

> . . . it's possible for an undetermined action not to be up to the agent who performs it. Just as supposing that an action is undetermined is no guarantee that it's up to the

[55] Clarke names an agent's exercise of freedom 'direct active control'.
[56] For more on causal deviancy, see §2.2.2 and §4.3.1.

> agent, so supposing that an undetermined event is someone's action is no guarantee that its occurrence is up to that person. In short, there is a "logical gap" between the premise that an agent-involving event is undetermined and the conclusion that it's up to the agent whether it occurs, a gap that isn't closed by adding that the event is something the agent *does*. [Shabo 2013, p. 154]

Shabo bases his argument on a variation of van Inwagen's famous 'replay argument' [2000], that we shall therefore briefly consider. Van Inwagen asks us to imagine an agent, Alice, faced by a choice: e.g., to to tell a lie or to tell the truth. Alice freely decides to tell the truth, which according to the libertarian implies that her act was undetermined. Imagine further that 'immediately after Alice told the truth, God caused the universe to revert to precisely its state one minute before Alice told the truth . . . and then let things "go forward again."' [van Inwagen 2000, p. 14] Since her original act was undetermined, Alice may now one again tell the truth or lie. Van Inwagen continues:

> Now let us suppose that God a thousand times caused the universe to revert to exactly the state it was in [. . .]. As the number of "replays" increases, we observers shall—almost certainly—observe the ratio of the outcome "truth" to the outcome "lie" settling down to, converging on, some value [. . .]. If, after one hundred replays, Alice has told the truth fifty-three times and has lied forty-eight times, we'd begin strongly to suspect that the figures after a thousand replays would look something like this: Alice has told the truth four hundred and ninety-three times and has lied five hundred and eight times. Let us suppose that these are indeed the figures after a thousand replays. Is it not true that as we watch the number of replays increase, we shall become convinced that what will happen in the next replay is a matter of chance? [van Inwagen 2000, pp. 14-15]

And since we will suppose that what happens in the 1001st replay is a matter of chance, we will equally suppose that Alice's decision in the original scenario—and in fact, in all the replays—was merely due to chance: each decision had an 'objective, "ground floor" probability' equal to the number of replays in which it occurs divided by the total number of replays.[57] Van Inwagen argues from this that Alice was not free—and many have agreed.[58] For he claims that Alice, in each single replay, lacks

[57] Recently, Buchak [2013] has argued that the rollback argument is unsound. She rightly points out that from the fact that after any large number of roll backs, there will always be some number of truth-tellings x, and some number of lyings y, it does not follow that there was an x percent objective probability that Alice would tell the truth. This application of the law of large numbers is only warranted if we already know that the results in each roll back are due to a stochastic process. But of course, the event-causal libertarian has already granted that the relevant processes are stochastic: on the event-causal libertarian picture, each action the agent considers has an objective probability of occurring. So strangely, the replay argument seems to be a mistaken argument for a thesis that Van Inwagen's opponent has already granted.

[58] Although van Inwagen aims to demonstrate that agent causation is irrelevant to free will, it seems irrelevant to the argument whether we suppose that Alice agent-causes her decisions, or that her decisions are indeterministically event-caused by her reasons.

an ability relevant to free will. According to Shabo, we should understand the ability that Alice lacks as 'a power *over* whether she voluntarily and intentionally lies or voluntarily and intentionally tells the truth' [Shabo 2011, p. 299]. She may have both the ability to lie and the ability to tell the truth (i.e., the real possibility of intentionally lying and of telling the truth), but she lacks the ability to determine which of these two will occur:

> . . . in virtue of what does Alice enjoy the power to make it the case that one of the two appropriately caused, reason-based decisions occurs, to the exclusion of the other, equally probable decision? [Shabo 2011, p. 299]

Again, Shabo claims that 'we can consistently allow that the outcome is settled *when* Alice makes up her mind to tell the truth [. . .] while denying that it's up to her *how* the outcome is settled.' But all this is but a tiresome repetition of moves. At the end of the day, Shabo has nothing to prevent the libertarian from defining the power to settle as follows:

> . . . an agent S exercises the power to settle in performing an action φ at t if and only if he exercises his ability to φ at t and had the ability to ψ at t [Franklin 2012, p. 397]

One's control over φ'ing-instead-of-ψ'ing is realized *by* choosing to either φ or ψ. That nothing *before* the agent's decision settles which outcome will ensue can hardly be an argument against the libertarian: the replay argument 'simply *describes* libertarianism in a rather colorful way' [Franklin 2012, p. 409]. Compare the contrastive formulation of the luck objection. When Mele asks why two possible worlds suddenly diverge at the moment of an agent's decision, the libertarian replies:

> They diverge when I exercised my free will. That's what matters. When you act freely, you exercise control when you act. [Clarke 2005, p. 417]

The libertarian thus seems to have a cheap reply to the luck objection: it simply fails to take regard of what an agent's exercising rational control *is*, on the very same theory of action (CTA) that the libertarian's critic espouses.

1.4.1 What good is indeterminism?

The reply described above essentially defuses the luck objection by pointing out that the objector's claim that an undetermined action cannot be under the agent's rational control directly contradict his (the compatibilist's) own views about what constitutes such rational control. The reply thus downplays the differences between a picture on which the causal connection between reasons and action is indeterministic and one on which it is deterministic: in either case, an agent exercises his rational control

when an action or decision is caused—no matter whether something else *could* have happened before this happening.

Effective as it might be against the luck objection, as posed by compatibilist or hard incompatibilist adherents to CTA, this reply leads to a new problem for the libertarian. For if the reason why an indeterministically caused action is not out of the agent's rational control is simply that it is caused and rationalized by her reasons, a real question emerges about the *necessity* of indeterminism for freedom. If event causation by reasons is enough to secure rational control, why should deterministic causation be incompatible with freedom? If causation and rationalization by reason states is enough to confer the agent's control on the result of an indeterministic causal struggle in her brain, why could it not confer it on the result of a deterministic causal struggle?

The libertarian may wish to reply that what is gained by the presence of indeterminism is precisely 'freedom of the will': rationalization and causation by reason states, be it deterministic or indeterministic, is enough to confer rational control on an action, but *freedom* is only gained if the action is undetermined. As Shabo claimed, there must be a 'logical gap' between an event's being an action and it's being 'up to the agent'.

But by reopening that logical gap, the libertarian would make himself vulnerable to the luck objection once more. If the libertarian admits that there *is* a type of control that is not secured by rationalization and event causation by mental states, the door is again open to the claim that *that* kind of control is incompatible with luck—*unless* there is reason to suppose that this further kind of control is not only compatible with indeterminism (for in that case, still nothing will be gained), but positively requires it. So the libertarian must give a positive characterization of the kind of control that is gained by indeterministic causation. Let us look at some proposed solutions in order to appreciate the difficulty of achieving this.

Franklin has argued that the answer to 'the problem of enhanced control', as he has styled the present challenge to libertarianism, is that the kind of control that is provided by indeterministic (and not by deterministic) event causation is characterized by having multiple *opportunities*:

> it is often mistakenly assumed that an agent's control is wholly exhausted by the agent's powers and abilities. I argue, however, that control is constituted not just by what we have the ability to do, but also by what we have the opportunity to do. [Franklin 2011b, p. 687]

He argues that deterministic agents do not have the opportunity to act otherwise than they do. This is obviously correct, but only because it is just another way to state the thesis of determinism. But what the libertarian needs is a *reason* to insist on

the presence of indeterminism—not another 'colorful *description*', to use Franklin's own words (see §1.4), of her position. Now it is striking to see that Franklin has *nothing* by way of argument for the thesis that multiple real opportunities adds to an agent's freedom. He says:

> The opportunity to do otherwise is not simply another opportunity on top of the many opportunities that compatibilist agents already possess. Rather, it is a significant addition. It affords agents with the opportunity to direct their lives in more than one way, to author how their lives unfold, and to choose from among several causally open options, thereby taking a stand on the kind of person they will become. This is no trivial addition. Indeterminism, therefore, is relevant to enhancing control because its existence is necessary for agents to possess the freedom to do otherwise. [Franklin 2011b, p. 704]

So indeterminism is a necessary requirement of free will because its existence is necessary for agents to possess the freedom to do otherwise—but 'the freedom to do otherwise' just *is* the real possibility to do otherwise, and that just *is* the presence of indeterminism. Franklin says that the freedom to do otherwise is 'significant' because without it, agents do not have the opportunity to, e.g., 'choose from among several causally open options'—but again, that is obviously circular. The rhetorical evocation of the significance of 'taking a stand on the kind of person [an agent] will become' is hollow—the compatibilist will agree, but simply give a deterministic interpretation. Similarly, Kane, Pereboom, and others propose that the reason why free will is incompatible with determinism is that freedom requires us to be the 'ultimate source' of our actions. But then 'ultimate source' is *defined* precisely as the libertarian demand for indeterminism, as so cannot be a ground for adopting it.

Clarke [2003, p. 93] offers his agent-causal account precisely as a solution to the present challenge. What makes indeterminism a necessary ingredient of freedom, he claims, is that it allows for actions to be agent-caused. But as we have already seen (§1.3.2), agent causation does not by itself make an event into a free action—if inanimate objects can also be substance causes, this kind of causation obviously adds nothing. Moreover, as Clarke [2003, pp. 193-196] himself admits, there is no reason to suppose that substance causation could not be deterministic.[59] So even *if* substance causation is necessary for freedom, why should it involve indeterminism? Even if agent causation could conceivably be indeterministic, Clarke simply does not answer the question what this would secure by way of freedom.

Finally, consider Mele's [2006] answer to the question on what grounds the libertarian may insist on indeterminism. He goes to great lengths to argue, along the same lines as above, that compatibilism may secure the same kind of rational con-

[59] And Markosian [1999] has even defended the possibility of deterministic agent-causal free will.

trol as libertarianism. In fact, according to Mele's 'soft libertarian' position (which he develops as a response to the luck objection), *most* of our actions are probably deterministically caused. But nevertheless, he claims, we may *value* the existence of indeterminism in our actions:

> Unlike hard libertarians, soft libertarians leave it open that determinism is compatible with our actions' being up to us in a way conducive to freedom and moral responsibility. However, they believe that a more desirable freedom and moral responsibility require that our actions not be parts of the unfolding of deterministic chains of events that were in progress even before we were born. If soft libertarians can view themselves as making some choices or decisions that are not deterministically caused [...] then they can view themselves as initiating some causal processes that are not intermediate links in a long deterministic causal chain extending back near the big bang. [Mele 2006, p. 97]

Again, the circularity of Mele's reason for why indeterminism may be 'more desirable' is obvious. And it is a rather small circle, too: if an agent's action is undetermined, that is desirable because the agent can view her action as... undetermined (a 'causal process that is not an intermediate link in a deterministic causal chain'). Compromising views like that of Mele and Kane—who also claims that deterministically caused actions can be free, as long as they trace back to *some* undetermined 'self-forming actions'—clearly illustrate the cost of avoiding the luck objection by downplaying the role of indeterminism: indeterminism becomes utterly unimportant.

1.5 Towards a reasonable libertarianism: a hypothesis

We have seen how the luck objection, for all its intuitive power, cannot be maintained once we realize that the event-causal libertarian has at his disposal the same account of rational control as others who espouse CTA. However, this realization has led to an embarrassing new problem: indeterminism does not offer anything that was not already secured on deterministic CTA. Whenever a libertarian tries to explicate why *freedom*, rather than just the rational control also available to a compatibilist, requires indeterminism, a combination of hollow rhetoric and colorful redescriptions of libertarianism follows.

The libertarian's (compatibilist) critics might think this is good news. But the nature of the defeat ought to worry all participants in the debate. For it seems that the libertarian has not even managed to *formulate* his position: he has not managed to state *for what* indeterminism should be necessary. That is bad news to all. For as Wiggins says, 'perhaps libertarianism is in the last analysis untenable', but:

> Compatibilist resolutions to the problem of freedom will always wear an appearance of

> superficiality, until what they offer by way of freedom can be compared with something else, whether actual or possible or only seemingly imaginable, that is known to be the best that any indeterminist or libertarian could describe. [Wiggins 1987b, p. 270]

Even those who are not embarked on the project of defending the libertarian view should therefore take it as a cause for concern that the libertarian position seems so close to collapsing into compatibilism. For, I suggest, this is a clear sign that something has gone terribly wrong in the contemporary debate. I will now attempt to identify the flaw underlying the debate, and offer an alternative route for the development of 'a reasonable libertarianism' [Wiggins 1987b]—a hypothesis that I will attempt to develop and establish in the coming chapters.

Let us make the dialectical situation that has led to the collapse of libertarianism explicit. As we have seen, the libertarian is forced to accept the following, on pain of succumbing to the luck objection:

> (**No Gap**) There is no logical gap between some event being an intentional action and its being up to the agent.[60]

For suppose there would be such a gap, as e.g. Shabo [2013] argued. Then there must be some form of control or 'up-to-usness' *other* than rational, or intentional control. And as we have seen (§1.4.1), it seems impossible to say what that would consist in. The libertarian must thus accept that an action's being 'up to the agent' (i.e. being free), is fully accounted for by her exercising rational control over it (i.e., by its being an intentional action). But now the libertarian also adopts CTA, from which it follows that:

> (**Neutrality**) An event's being an intentional action does not depend on whether that event was physically determined.

Adopting CTA, and thus (Neutrality), ensures that, *pace* the luck objection, undetermined events *can* be 'up to the agent'. But once we combine (No Gap) and (Neutrality), it becomes impossible to insist that freedom requires indeterminism—for then an action's being up to the agent is neutral with regards to determinism or indeterminism. From this follows the position that I have labeled neutral compatibilism:

> (**Neutral Compatibilism**) Freedom of the will is compatible with both determinism and indeterminism.

So it appears that, once she has accepted CTA, the game is over for the libertarian. Should we conclude that libertarianism is an indefensible position? I think not.

[60] Excepting, perhaps, in the sense in which a coerced intentional action is not up to the agent. But the libertarian should grant that a coerced action can nevertheless be an exercise of free will: the libertarian's position is precisely that, even if there are many factors which rationally pull us in a certain direction, we are still from to perform or not perform that action.

Rather, I suggest that we should take seriously the idea that *CTA is an inherently compatibilist doctrine*. That would straightforwardly explain why, once we adopt CTA, it becomes impossible to formulate a plausible libertarian position. For then contemporary libertarians are trying to defend the necessity of indeterminism after having, in effect, already bought into neutral compatibilism. Moreover, it would explain why those who raise the luck objection have so little success: by embracing CTA, these critics have equally already accepted neutral compatibilism. And of course, it also explains why libertarians do not succeed in pressing the Consequence Argument against compatibilism.[61]

Interestingly, the originator of CTA seems to agree. Davidson proposes to analyze '*A* is free to do *x*' as 'he would do *x* intentionally if he had attitudes that rationalized his doing *x*' [Davidson 1973, p. 79]. That is, he claims that acting freely *is* what is, on his account, acting intentionally.[62] Davidson may appear not to be a neutral compatibilist: he claims that free actions *are* causally determined. But this is merely due to a peculiarity of Davidson's account of causation. Causation, for him, is subsumption under exceptionless generalizations, and hence he believes it can only be deterministic [Davidson 1967]. However, there is nothing in Davidson's account to object to the idea that, *if* there were indeterministic causation, free actions might be undetermined.

If it is true that to accept CTA already is to commit oneself to compatibilism, I think it is clear what the libertarian's response should be. Rather than falling into the trap of trying to show how indeterminism 'enhances' the control that an agent exercises on CTA, she should reject CTA as an altogether mistaken account of intentionality. Instead, the libertarian should defend a theory of intentional action that intrinsically requires indeterminism. That is how she can accept (No Gap) while rejecting (Neutrality). Libertarianism is, or should be, the position that *acting itself* is incompatible with determinism.[63]

[61] As I noted in fn. 32, the Consequence Argument is best seen as an argument against the possibility of choice, and hence intentional action, under determinism. If that is right, then it is impossible to accept both CTA *and* the Consequence Argument.

[62] Davidson then goes on to say that this cannot be a fully sound analysis of '*A* is free to do *x*', because instances of so called deviant causal chains provide counterexamples to the claim that an action is intentional iff they are caused and rationalized by his mental states. But the analysis is only flawed because the conditional does not give a sufficient condition for acting intentionally. Davidson claims that this problem should nevertheless not deter us from saying that freedom to act is an agent's possession of certain causal powers (his mental states)—we just cannot give the precise empirical conditions that should be added to the conditional 'if he had attitudes that rationalized his doing *x*'. See §2.2.2 and §4.3.1 for more on causal deviance.

[63] Bishop [1989, p. 59] agrees: 'Incompatibilists would do well to rally around the slogan that to act is one thing, to be caused to act by one's own mental states, quite another. All the doubts people continue to have about how action could possibly consist purely in behavior with suitable mental causes can be brought to bear on the incompatibilist side of the debate.' Of course, Bishop himself accepts CTA, and is a compatibilist for that reason.

This proposal constitutes a radically different approach from what is common in the contemporary debate. For as we have seen in this chapter, the question of the free will debate is almost always taken to be: what is necessary for 'transforming a mere action into a free action' [Franklin 2011a, p. 203]? Compatibilists and libertarians alike answer by giving certain criteria—e.g., lack of coercion for the compatibilist, indeterminism for the libertarian.[64] And the assumption that free action is action *simpliciter* that conforms to certain extra conditions may seem innocuous. How else to understand 'free' in 'free action' or 'free will' other than as an adjective modifying an independently understandable concept? However, we should note that neutral compatibilists like Davidson and Bishop [1989] do not share this assumption. For them, acting for a reason is intrinsically free action. The libertarian, I suggest, should argue that it is this freedom *in* action—which I have identified with the *spontaneity* of acting for a reason in the introduction—which requires indeterminism.

The thought that the distinction of compatibilism or incompatibilism is applicable to a conception of acting for a reason itself is so alien to contemporary philosophy precisely because of the prevalence of CTA. We are used to thinking about 'decisions' and 'intentions' as mental states or even brain states with certain content, and it seems to be an entirely contingent matter whether the connections between such states and the actions they cause are deterministic or indeterministic. To make room for the kind of libertarianism I have in mind, we must learn to see this way of thinking for what it is: a set of assumptions forced on us by a reductive philosophical outlook. In the next chapter, I will try to get this reductive outlook in focus, and argue that it is optional. We do not *need* to accept CTA. Moreover, I will also explain how the libertarian can evade the luck objection if she is to reject CTA.

Of course, that will not be enough to defend libertarianism. To do that, I will have to show that CTA is a false doctrine. I begin that task in §2.2.2, but will only finish it in §4.3.1. And finally, I will have to show why acting for a reason positively requires indeterminism. That will be the challenge I take on in chapters 3-5. There is thus much work to do in order to defend a reasonable libertarianism. But after witnessing the hollow rhetoric and repetition of moves characteristic of the contemporary discussion, I suggest that, if there is any interesting question about free action at all, it must be the question whether intentional action is possible under determinism.

* * *

[64] An exception to this is formed by compatibilist positions like Davidson's (discussed above), and Steward's [2012] agency-incompatibilism, which I will discuss in §5.3.

Chapter 2

Accidentality in thought and action

We have seen that, once we accept CTA, the position that I labeled *neutral compatibilism* becomes unavoidable. I suggested, therefore, that in formulating a plausible libertarian conception of free will, we ought to abandon CTA and develop an account of intentional action as already requiring indeterminism. On such an account, then, acting intentionally would not be for an agent to be caused to do something by some of her psychological states. But at the same time, such an account would have to evade the luck objection: it must not be a mere *accident* that the agent acts as she does. However, it is not yet clear how these two requirements can be combined. What could it mean to say that an action is undetermined, not caused (as on the event-causal model), yet not a matter of luck? In this chapter, I set out to investigate this.

I will argue that our puzzlement about this question is grounded in a certain picture of what it means for something to be a matter of luck (§2.1). According to this picture, which I call the *uniform conception of luck*, an event is (roughly) lucky if it does not occur in a wide range of possible worlds close to the actual world. I investigate this conception by comparing discussions of the luck objection with the debate on *epistemic* luck. I then argue that this conception of what it is for a belief or action to be accidental fails to achieve its purpose: it cannot explain why accidentality is incompatible with knowledge and freedom, respectively. Furthermore, I argue that the uniform conception is implicitly based on a broadly Humean understanding of modality: it starts from a philosophical picture on which there are *no* non-accidental

connections, and then tries to re-import a semblance of non-accidentality by means of possible worlds talk.

Instead, I propose that we should look at non-accidentality in a different way: for something to be non-accidental is for there to be an *explanation* of why it happens. And, as there may be different *forms* of explanation, there may equally be different forms of non-accidentality. One of these, I argue in §2.1.2, is for an event to have a *reasons-explanation*, i.e., an explanation in virtue of which it is an intentional action. And, I suggest, for an action to be free just is for it to be non-accidental in that sense. Free action just is intentional action.

Now, CTA offers us a reductive account of the form of explanation that constitutes intentional action: it says that the relevant kind of explanation just is the *same* that we find everywhere else in nature—event causation is the *only* kind of non-accidentality. Once we accept this reduction, compatibilism becomes inescapable. But, I argue (§2.2.1), there is no *prima facie* reason to accept the reduction. And although I will not yet offer a full-blown argument against CTA at this point (I will do that in §§4.2.3-4.3.1), I suggest that the well-known problem of deviant-causal chains shows that CTA may well be unable to account for the non-accidentality required for intentionality (§2.2.2). I conclude that we should take seriously the hypothesis that reasons-explanation may be a *sui generis* form of non-accidentality, or (what I explain to be the same) a form of causality. And if that is right, then *in*compatibilism becomes an option again: it may be that reasons-explanation excludes physical determination.

2.1 Forms of accidentality and forms of explanation

In the previous chapter, we have seen that adopting CTA and hence neutral compatibilism makes it possible to evade the luck objection. But at the same time, it seems, there still is a kernel of truth to the idea that freedom is somehow opposed to arbitrariness or randomness. The intuitive thought remains: if something is a mere accident, it cannot be 'up to the agent'. Although the combination of CTA and neutral compatibilism may succeed in evading the luck objection, it does little to explain (or explain away) the plausibility of that idea. And indeed, it seems that proponents of event-causalism find themselves torn between two competing intuitions: on the one hand, they want to agree that accidentality and freedom are opposed, while on the other, they claim that free actions may, in principle, come about through probabilistic processes. For example, here is Franklin (the event-causal libertarian whose strategy for evading the luck objection we have discussed in §1.4):

> . . . van Inwagen is clearer than many other proponents of the luck argument concerning exactly what he means by chance: an event is a matter of chance just in case the

> objective probability of its occurring is less than 1. [...] From the fact that there is an objective probability of less than 1—specifically 0.5—that Alice will choose to tell the truth, van Inwagen concludes that 'in the strictest sense imaginable' the outcome is a matter of chance. If this indeed is what van Inwagen means by chance, then [the claim that undetermined actions are a matter of chance] is certainly true. All undetermined actions have an objective probability of less than 1 of occurring and so all their occurrences are a matter of chance. [Franklin 2011a, p. 216]

Thus Franklin admits that undetermined actions, on the event-causal picture, are 'in the strictest sense imaginable' a matter of chance.[65] If that is so (and if we want to allow for undetermined but free action), should we then ultimately reject the intuitive thought that accidentality threatens freedom? I think not—and neither, interestingly, does Franklin:

> If indeterminism entails luck and randomness, then indeterminism does indeed appear to be incompatible with free will. [...] [T]here seems to be an inverse relation between luck and control: the more an action is subject to luck, the less it is under our control, and the more an action is under our control, the less it is subject to luck. Luck and control thus appear to exclude each other: an action cannot be both wholly a matter of luck and wholly under our control. [Franklin 2011a, p. 200]

It therefore seems that, although all parties in the debate can agree that accidentality threatens freedom, we do not quite understand *how* it does so. To overcome this, I will provide a detailed account of the sense of accidentality that underlies contemporary formulations of the luck objection in the following section (§2.1.1). I will show that it is characteristic of this dominant conception of accidentality—which I will call the uniform conception of luck—that it can be defined without reference to the phenomena it is allegedly incompatible with—e.g., freedom. Luck, on this view, is nothing but metaphysical contingency. Once I have reconstructed this view, I argue that it is difficult to see how luck or accidentality, conceived in that way, could have anything to do with free action. Instead, I suggest in §2.1.2 that a different notion of accidentality underlies the original intuition that luck threatens freedom. With this improved understanding of the way in which accidentality and freedom are related in hand, I will argue that an undetermined action need not be a mere accident, and that we do not *need* to accept the compatibilist event-causal framework in order to evade the luck objection.

2.1.1 A uniform conception of luck

Exactly what is it for something to be a matter of luck? And what is it about luck that, at least supposedly, undermines freedom? In this section, I will reconstruct

[65] Franklin is here using chance as a synonym for luck.

the answers given to these questions in the contemporary debate. And as we will shortly see, that is not just the debate about free will. For the conception of luck at play in the discussion of the luck objection is one that is imported from other areas of philosophy—especially from recent discussion in epistemology. After explicating the relevant conception, I will argue that it does not in fact do justice to the original intuition that luck undermines freedom, and will suggest a different way of spelling out that thought.

To begin this reconstruction, let us first note that there is a strong intuition that luck is inimical to an agent's control over certain outcomes. If it is true that, e.g., it is just a matter of luck that a marksman hits his target, we want to say that the result was not (fully) due to his expertise as an archer, and we feel that the outcome was not under his control. Similarly, nobody controls the outcome of a fair lottery, precisely because it is just a matter of chance what the lucky numbers turn out to be. This can make it seem as if there is a phenomenon of 'luck', pure and simple, which, whenever it occurs between an agent's attempt to get something right and the resultant outcome, destroys the agent's claim to say that the result was his or her *achievement*. The task of the philosopher, then, is to explicate what it is for this phenomenon to obtain. How can this be done?

An outcome (the arrow hitting the target, or the agent's deciding to φ) is a matter of luck, intuitively, if it might as well *not* have occurred, given the same initial conditions (an agent's skill at archery, or the balance of her reasons)—we say, e.g., that the archer might as well, or might easily, have missed the target. So this seems like an excellent opportunity to utilize one of the modern analytic philosopher's favorite tools: the framework of *possible worlds*. And indeed, such is the epistemologist Duncan Pritchard's definition of luck:

> If an event is lucky, then it is an event that occurs in the actual world but which does not occur in a wide class of the nearest possible worlds where the relevant initial conditions for that event are the same as in the actual world.[66] [Pritchard 2005, p. 128]

Luck, then, is a phenomenon that applies in the first instance to *events*, of whatever kind. So this account of luck can potentially be used to elucidate a host of philosophical notions which intuitively have some relation to luck—and indeed, it *has* been applied in a variety of domains in philosophy.[67] In order to understand the importance of this conception of luck for the debate on the luck objection, and where it goes wrong, it will be instructive to first consider one of these other domains. I will

[66] Note that for Pritchard, this is only a partial definition of luck. According to him, another aspect of luck is that the relevant event must be significant for the agent, so that the agent can have good or bad luck. I will ignore this detail in what follows.

[67] As we will see below, most prominently in the debates on epistemology and free will, but arguably the same notion of luck is also at play in the debate on so-called moral luck. See for example Peels [2015].

take as my example the phenomenon which is Pritchard's main interest: *epistemic luck*. As we will see, there is an interesting analogy between the role that luck plays in the contemporary epistemological discussion and the debate on freedom.[68]

Many contemporary epistemologists attempt to define *knowledge*. A naive definition says that knowledge is justified true belief. That definition is confronted with counterexamples known as the so-called Gettier cases: e.g., a subject comes to truly believe that it is noon by looking at a clock that has in fact stopped ticking a day ago.[69] Clock-reading is a good method for telling the time, and it actually happens to be the time indicated by the clock. Yet we want to say that our subject does not *know* that it is noon. Why? Because, we feel pressed to say, *it is just a matter of luck* that our subject forms a true belief by looking at the clock. That, at least, is Pritchard's diagnosis. And as we will see, I have no quarrel with this: it *is* true that a Gettier-subject lacks knowledge because her belief is only accidentally true. However, I will suggest that Pritchard's conception of luck does not capture the sense in which this is true.

On the basis of his diagnosis of these cases, Pritchard proposes to amend the justified true belief definition of knowledge by adding a fourth condition: it must not be a matter of luck that the subject acquires the true belief.[70] All that remains is to give a precise formulation of this 'no luck' clause. This is a straightforward exercise of applying the general formula of luck to the case of belief-formation:

> If it is a matter of epistemic luck that S forms a true belief p based on evidence q, then in a wide class of the nearest possible worlds in which S forms the belief that p based on the evidence that q, p is false.[71]

Thus it seems reflection on Gettier cases teaches us something substantive about the nature of knowledge: we must include a 'no luck' clause in the definition of knowledge. It may be that knowledge is (1) a belief that is (2) justified, (3) true, and (4) not a matter of epistemic luck. Or alternatively, the anti-luck condition may even supplant the justification condition.[72] Either way, the subject who comes to acquire the belief that it is noon by looking at the broken clock does not *know* that it

[68] Levy [2011] notices this analogy and explicitly adopts Pritchard's definition of luck in his attempt to formulate a successful version of the luck objection.

[69] This example is originally due to Russell [2009, p. 91]. It is thus not one of Gettier's [1963] original cases, but is still widely taken as a Gettier-style counterexample to the 'justified true belief' definition.

[70] This move is of course not unique to Pritchard. The idea that the Gettier cases show that the naive definition suffers from a problem of luck is widespread, as is the idea to amend the definition by including a so-called 'safety' condition. See Pritchard [2005, pp. 145-152] for an overview. I am using Pritchard as a paradigmatic example of this tendency because of the exceptional clarity with which he illustrates the conception of luck I am reconstructing.

[71] This is a more detailed rendering of the definition Pritchard [2005, p. 146] gives. To be more precise, the kind of epistemic luck that Pritchard here defines is *veretic* luck, as opposed to *reflective* luck. The latter plays a role in his analysis of the skeptic's position, but not in that of the Gettier cases.

[72] Pritchard [2005, p. 173] claims to be neutral about whether or not his account of epistemic luck requires justification as an additional requirement. According to him, that issue must be decided in the broader debate between internalism and externalism.

is noon, because the anti-luck condition is not satisfied: in a wide range of possible worlds in which she acquires the same belief based on the same evidence, her belief is false—e.g., in those worlds in which she looks a bit later or a bit earlier.

Here we find an interesting parallel between cases of epistemic luck and luck in action. Epistemic luck destroys a subject's claim to knowledge—that she reached the conclusion by able exercise of her epistemic abilities. Just so, it seems, luck in action supposedly (according to proponents of the luck objection) destroys an agent's freedom—her claim that the action was an exercise of her practical deliberative abilities. Such luck in action, on the picture we are considering, is then to be understood in the same way as epistemic luck: the performance of a certain action by an agent is a matter of luck if there is a wide class of close possible worlds (i.e., worlds in which the agent has sufficiently similar preferences, beliefs, etc.) and the action does not occur. We can see that it is indeed the same conception of luck that is at play in both the discussion on epistemic luck and the luck objection by comparing Mele's formulation of the latter:

> . . . [an agent's] making that alternative decision rather than deciding in accordance with his best judgment—that is, the difference between [possible world] *W* and the actual world—is just a matter of bad luck or, more precisely, of worse luck in *W* for the agent than in the actual world. After all, because the worlds do not diverge before the agent decides, there is no difference in them to account for the difference in decisions. [Mele 2006, p. 8]

An action's being a matter of luck is just for it to happen in some possible worlds with initial conditions that are sufficiently similar to the initial conditions in the actual world. On this definition, it is clear that any action conforming to (Incompatibilism), as I formulated it in §1.1, is a matter of luck. For according to the libertarian, it is possible for any action not to occur given exactly *the same* initial conditions, and so the relevant possible worlds will thus be very close indeed. So undetermined actions will be lucky and hence unfree.[73]

I think we can thus see that epistemic luck and luck in action, on the Pritchard/Mele picture, are simply instantiations of the same general phenomenon of *luck*, pure and simple. In both cases, of epistemic luck and of luck-in-action, what it means for the relevant events to be lucky is the same—the verbal distinction between 'epistemic' and other forms of luck is based merely on the kinds of events and initial conditions we are interested in.[74] And in both cases, the idea seems to be that a sound analysis of

[73] Mele takes it that, at least for the case of luck in action, 'close enough' must mean that the initial conditions are exactly the same as in the actual world—such luck can only exist given indeterminism. As I said in fn. 75, Levy disagrees. According to him, an action is lucky even when there are close possible worlds in which the outcome does not occur given only *slightly different* initial conditions.

[74] And of course, the difference in subject matter may allow for differences in what it means for initial

a certain target phenomenon—knowledge and free will, respectively—requires a 'no luck' clause. For just as Pritchard proposes to include an anti-luck condition in the definition of knowledge, so it seems the proponent of the luck objection is insisting that a sound definition of free action must include a similar 'no luck' clause: free action is (1′) intentional action that (2′) is not a matter of luck (possibly with some additional requirements).[75]

It thus seems that the phenomenon of luck can crop up in many places. The conception of luck under investigation here is thus a *uniform* conception: there is but one sense in which something can be a matter of luck. Crucially, this picture thus assumes that luck is something we can understand independently from the concepts that we are investigating—knowledge on the one hand, and free action on the other. If it were not, it could not be part of a set of individually necessary and jointly sufficient conditions that make up a definition of the target phenomenon. The picture we get is one of luck as a force that robs agents of control over an outcome wherever it rears its head. But is luck, on the uniform conception, able to play the role envisioned for it? I will now argue that it is not.

Consider first the epistemic case. If 'no luck' is a fourth element of the definition of knowledge, then such luck must be understandable without reference to the concept of knowledge. Indeed, it is precisely the aim of many contemporary epistemologists to tell us what knowledge *is* in terms of certain more basic and presumably more easily understandable elements. That is why Pritchard employs the uniform conception. However, it seems that the very uniformity of luck risks making it utterly unclear why luck ought to figure in a definition of knowledge at all. For consider the dialectics that lead Pritchard to adopt the 'no luck' condition. We noticed first that there are cases, such as the broken-clock scenario, in which it seems obvious that a subject lacks knowledge. We then discovered that we can describe all (or at least many) of these as cases in which, holding certain conditions fixed, there is a wide range of close possible worlds in which the subject's belief is false. But how does any of that amount to an explanation of *why* being in such a position is incompatible with knowledge?

Suppose that Pritchard answers along the following lines: the reason why the nearness of possible worlds in which the subject's belief is false is incompatible with knowledge is that subjects who are in that predicament are *lucky*—and, as we all know, luck is incompatible with such things as knowledge, freedom, skill, etc. Would

conditions in a possible world to be close enough to those in the actual world: for example, epistemic luck may obtain even if the actual world is deterministic.

[75] One way to understand the hard incompatibilist is as saying that a free action would have to satisfy both the mentioned conditions, as well as (3′) the action must not be physically determined. The hard incompatibilist thus argues that the concept of free will is contradictory: nothing could satisfy it.

this answer not have an air of circularity to it? If it is true that luck undermines knowledge, then it seems that an *account* of luck would have to show how it can do this. We cannot just give a definition of luck in counterfactual terms, and then invoke the intuitive principle that luck, in general, undermines achievement. To paraphrase Lewis [1983, p. 366]: a man does not have a mighty biceps just in virtue of being called 'Armstrong'. Just so, it seems, the fact that we *call* someone who has certain counterparts in close possible worlds 'lucky' does not imply that she cannot have knowledge.

Of course, the definition of luck in terms of the counterfactual closeness of a certain outcome was not arrived at just by stipulation. As we have seen, the idea behind it was that an outcome is lucky when it might easily have been otherwise, as in, e.g., a lottery. Needless to say, the outcome of a lottery *is* a prime example of luck or randomness. However, the assumption that *all* luck, including the sense of 'luck' that is intuitively at odds with knowledge, must be analyzed in the same terms is precisely what is at issue here. And Pritchard does not seem to offer us any explanation of what is epistemically *wrong* with being in the predicament he describes as having epistemic luck. Now he might want to say: it is, epistemically speaking, bad to be in that predicament because a subject who is in it is *ipso facto* in a Gettier-case. But even if we grant this,[76] it does not show why luck and knowledge are incompatible. To borrow Franklin's [Franklin 2012, p. 409] words, saying that the subject has counterparts in close possible worlds with false beliefs is just a colorful redescription of her predicament: it is just to say that she is in a Gettier case.

So even *if* the definition of knowledge in terms of (1)-(4)[77] is extensionally correct, that does not show that having epistemic luck is the *reason* why Gettier-subjects lack knowledge. Pritchard claims that in Gettier cases, 'it is [the] counterfactual nearness of error that gives us the sense that the agent's true belief is just too lucky to count as knowledge' [Pritchard 2005, p. 158]. But this is questionable. In general, it seems, there is not a tight link between the counterfactual nearness of error and knowledge. When someone solves a difficult logic puzzle while the neighbors are playing loud guitar music, error might be counterfactually very near, but this does nothing to make us doubt that the agent knows the answer, *if* she manages to solve the puzzle. Of course, this example is not exactly analogous to cases of epistemic

[76] In fact, it is arguable that the anti-luck condition does not exclude all Gettier-style cases. Levy [2011, pp. 26-29] provides some reasons to think so. His argument centers around cases of overdetermination. It may be that a subject believes that p, p is true in many nearby possible worlds because many agents individually seek to ensure that p, and the subject believes that p because she reads a fake news report which mentions that p. Levy also argues that obvious modifications to Pritchard's account will not suffice. This strengthens the point that there seems to be nothing about the uniform conception of luck that explains why luck threatens knowledge, or achievement more widely.

[77] This assumes that we want to retain (3), the justification requirement, but see fn. 72.

luck as Pritchard defines them: in the nearby possible worlds, the agent doesn't solve the puzzle correctly. Still, it ought to make us wonder why the counterfactual nearness of error in Gettier cases should be any different. Is it truly this counterfactual nearness which undermines the status of a Gettier-subject's belief as knowledge? Or is that counterfactual nearness, perhaps, a consequence or *symptom* of the fact that something else is wrong—i.e., that she is not employing her epistemic abilities properly?[78]

To clarify: all this is not to deny that subjects in a Gettier case suffer from epistemic luck, and that they lack knowledge because of that. Rather, it is to question whether the uniform conception gives us an intelligible account of what it means for a belief to be accidentally true. Interestingly for this purpose, Pritchard has later argued that the anti-luck condition does not by itself provide an adequate account of when a belief can count as knowledge. He argues that there are cases in which (1)-(4) obtain, and error is thus *not* counterfactually close, in which the subject's belief intuitively still does not qualify as knowledge. On the basis of such cases, he proposes to add yet a further condition to the definition of knowledge: roughly, the belief must be the product of the subject's cognitive abilities. Now as Pritchard admits, there is *prima facie* a strong link between the idea that luck undermines knowledge, and the idea that knowledge requires that a belief be the result of a subject's exercise of her cognitive abilities:

> What does it take to ensure that one's cognitive success is not due to luck? Well, intuitively anyway, that it is the product of one's cognitive ability. Conversely, insofar as one's cognitive success is the product of one's cognitive ability, then—again, intuitively—one would expect it to thereby be immune to knowledge-undermining luck. [Pritchard 2012, pp. 248-249]

However, Pritchard argues that the ability requirement and the no luck requirement must actually be distinct necessary conditions of the definition of knowledge. He argues so precisely because, e.g. in Gettier cases, the ability requirement seems to be satisfied while the no luck requirement is not, and in other cases, the no luck requirement is satisfied whereas the ability requirement is not.[79] But it is important to note that this is true only provided that we already assume that the 'master intuition' [Pritchard 2012, p. 247] that accidentality undermines knowledge must be cashed out in terms of the anti-luck condition, and thus in terms of the uniform conception of luck. If we do not assume that from the outset, we may view the

[78] Horst [2015, p. 15] argues similarly with regard to attempts to dispel luck in action by appeal to the uniform conception of luck (which I will investigate in §2.2.2).

[79] The details of these cases are too complicated to go into here. For our purposes this is unimportant, because as I explain below, I am willing to grant that neither the anti-luck condition nor the cognitive ability condition (when these are conceived as independently understandable conditions) suffice for knowledge.

fact that there are beliefs which satisfy the ability requirement, but not the no luck requirement, as exemplifying the problem I have illustrated above: there just does not seem to be anything about epistemic luck (on the uniform conception) which is incompatible with knowledge. That is: it may be that a *proper* account of the non-accidentality relevant for knowledge must satisfy the intuitive connection between cognitive ability and the absence of luck that Pritchard points out.

My aim here is not to argue conclusively against Pritchard's analysis of knowledge and the notions of cognitive ability and luck that he employs. However, it will be instructive to shortly indicate what the alternative picture I am hinting at would look like. On that picture, we do not take the intuitive idea that knowledge is non-accidentally true belief as an *analysis* of knowledge. Non-accidentality is not part of a set of individually necessary and jointly sufficient conditions. Instead, we view the statement 'knowledge is non-accidentally true belief' as a nominal definition which tells us that knowledge is a *form* of non-accidentality: for a subject's belief that p to count as knowledge *means* that it is no accident that she believes it. That is to say that the subject does not just *happen* to have the belief, but there is a certain *explanation* of why she believes it. And this explanation cannot be, e.g., that she believes that p because an evil neuro-scientist triggered a device in her brain. The explanation of why she believes that p must be of a particular kind: it must, at the same time, be an explanation of why it is *right* to believe p—it must be a *justification*. The subject does not just happen to believe that p—it is no accident that she believes it—in the sense that she believes it *because*, say, she believes that q (and q implies p). On this view, the formula 'knowledge is non-accidentally true belief' just means that the concept of knowledge is the concept of something that has such an explanation.[80]

But is that not just to return to the naive definition of knowledge as justified true belief? And did the Gettier cases not show that this definition is untenable? Not necessarily. I think the lesson to learn from reflections on epistemic luck is that a correct account of justification must be truth-involving, and thus, that a Gettier-subject is *not* justified in believing what she does.[81] In the stopped clock case it is

[80] Compare Sellars' famous claim: 'In characterizing an episode or state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says' [Sellars 2000, p. 76]. On Sellars' picture, the concepts of justification, belief, knowledge, and arguably even truth will form a very tight circle. For it seems that to believe something—to hold it true—*is* to place it in the space of reasons, i.e., to justify it. If that is right, then to understand one of the elements in the 'non-accidentally true belief' formula, one has to understand all of them. In chapter 3, I give an account of belief and justification which conforms to this idea.

[81] Compare Rödl [2007, p. 147n10]: 'the obvious conclusion [of reflection on Gettier cases] is not that knowledge cannot be defined as justified true belief, but that this definition deploys a concept of justification according to which justified beliefs cannot fail to be true.' This comes down to accepting a kind of disjunctivism about justification. Disjunctivism about perception is commonly understood as the idea that someone who thinks that p on the basis of misperception or hallucination, which may be qualitatively indistinguishable from a veridical perception, is not justified in thinking that p (whereas a veridical

not difficult to defend this, at least if we notice that looking at a broken clock is not a justified way of coming to learn the time.[82] On the picture I am sketching, a Gettier-subject's belief lacks an explanation which shows the belief to be true, and is thus accidental. This preserves the thought that Pritchard [2012, pp. 248-249] finds so *prima facie* plausible: that epistemic non-accidentality consists in the sound exercise of one's cognitive abilities.

Of course, none of this is meant as a precise account of knowledge. In order to develop the picture I have only hinted at here, we would have to understand the relevant kind of explanation—i.e., we would have to understand what it is for a subject to believe something on grounds which entitle her to believe it.[83] But what I hope to have shown is that there is a different, more fruitful way of thinking about what it means for something to be accidental. For luck, on the uniform conception, is just the counterfactual nearness of a different outcome. And it seems to me that this counterfactual nearness is just a *symptom* of the fact that (in the epistemic case) the agent's belief does not come about through the sound exercise of a cognitive ability—it is not *itself* the reason why the belief is accidental, in the sense relevant to knowledge.

The difference between these two ways of thinking about accidentality will become clearer when we remember the philosophical pedigree of the modal definition underlying the uniform conception of luck. The framework of possible worlds, counterparts, and closeness relations which it takes for granted was, of course, largely popularized by David Lewis. Lewis' ambition, as a neo-Humean, is to develop a metaphysics on which there are no necessary connections. According to him, 'all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another' [Lewis 1986, pp. ix-x]. That is: one matter of fact and then another, with absolutely no connections between them. The elements of this so-called 'Humean mosaic' are just there, without a reason or explanation of why the elements are ordered as they are. Lewis then develops the counterpart framework as a way to make sense of modal talk: to say that it is necessary that S would believe that p given X, e.g., is to say that in all possible worlds in which counterparts of S are

perception *does* provide justification). My suggestion can be seen as an extrapolation of this thought to justification in general. That is, I think, in line with what at least some disjunctivists intend [e.g. McDowell 2009].

[82] It seems that part of what stands in the way of this picture of justification is the prevalence of an overly strict connection between having a justification and being blameworthy. It is clear that the subject who forms a belief about the time by watching the broken clock does 'nothing wrong' in a certain sense—we cannot blame her for being mistaken. But that we can excuse her mistake does not show that she has not made a mistake in the first place. Whereas believing arguably implies being blameless, the opposite is not necessarily true.

[83] The account I give of *reasoning* (inferring one belief from another) in chapter 3 goes part of the way to showing this.

in conditions X, those counterparts come to believe that p. He thus reduces necessary connections to non-modal connections between actually existing states of affairs in different worlds. The counterpart framework is a way to talk about necessary connections between things which, strictly speaking, are fully accidental.

Proponents of the uniform conception of luck, although they may not share all of Lewis's Humean ambitions, do much the same thing. They ignore the actually existing explanatory connection between a subject's belief that p, and (say) her prior belief that q, and then ask whether or not is an accident that she believes that p. If we ignore the question what the subject's *actual* reason for forming that belief was, the only sense that is left for the question 'was her belief an accident?' is 'do her counterparts also form it?'. The uniform conception of luck assumes that the only relevant sense of accidentality is the counterfactual nearness of a different outcome. And it seems that whether or not a different outcome is counterfactually near, on that conception, must be a brute fact—just as on Lewis' picture. For if we *explain* the counterfactual nearness of a different outcome by, e.g., pointing out that the agent does not properly exercise her epistemic abilities, it will not be *luck* (in the uniform sense) which undermines the agent's claim to knowledge. This is why I think luck, on the uniform conception, is unable to play its intended role of undermining an agent's claim to control a certain outcome. When we want to know whether a belief counts as knowledge, or an action is free, we want to know what explains the actual occurrence of that event. And there *can* be explanations which are consistent with the counterfactual nearness of a different outcome—that is illustrated, e.g., by the possibility of indeterministic causal explanations.

On the alternative picture of (non-)accidentality that I favor, we thus take seriously the idea that there are explanatory connections between things in the actual world.[84] It is not a brute fact that, e.g., an agent first believes q and then p. Rather, the latter state or occurrence *depends on*, and hence is explained by, the former. Justification signifies a kind of explanatory, and thus modal, connection between happenings. From this perspective, including a 'no luck' clause in a definition of knowledge is somewhat akin to listing the cake on its own recipe. Or to be more precise, it is an attempt to define what it is for a belief to depend on, or be grounded in, something else—a form of non-accidentality—after first having eliminated all modal connections. If that is right, then we should not be surprised if counterexamples to definitions of knowledge that employ the uniform conception of luck will continue to crop up. At

[84] Of course, the neo-Humean will counter that what it is for one event in the actual world to explain (or cause) another is for certain states of affairs to obtain counterfactually. But I think that Lewis' mission statement, cited above, makes clear that this would be to explain away explanation (or causation)—at bottom, on his view, everything is just accidental. I argue again neo-Humean accounts of causation in more detail in §5.2.1.

the very least, I think the above provides reasons to reconsider our confidence in the possibility of giving such a definition.

2.1.2 Accidentality and intentionality

I have argued against the uniform conception of luck that there is nothing on that conception of luck or accidentality that can undermine an agent's claim that a certain outcome is 'up to her', in a broad sense. Until now, I have illustrated this by considering the case of epistemic luck. But, I argue, the uniform conception also fails to do justice to the original intuition underlying the luck objection: that freedom and accidentality are incompatible. For imagine the proponent of the luck objection against the possibility of undetermined but free action saying: 'I call an action lucky if it is metaphysically contingent, and since it is generally true that luck destroys achievement, any action which is metaphysically contingent cannot be free.' Just as in the epistemic case, that will not do. If it is true to say that luck destroys freedom, then a good account of such luck ought to show *how* it does so. Without an argument to that effect, the claim would remain hollow. It is not enough for the proponent of the luck objection to point out that we view luck as a negative phenomenon, and then to hope that this sense of unease transfers to anything that conforms to his definition of luck.

The uniform conception of luck thus arguably fails to make sense of the idea that accidentality threatens freedom. But there is a different way in which we can account for that intuitve thought. In the epistemic case, I suggested that a belief's being non-accidental consists in its having an explanation of a certain kind: a justification. If someone's belief has such an explanation, we credit her with knowledge. Knowledge *is* the concept of that form of dependence between one thought and another (or between a thought and, e.g., the capacity of perception) in that way. Accidentality is incompatible with knowledge because knowledge is *itself* a form of non-accidentality. Analogously, then, it would seem that if there is a sense of 'accidentality' that is incompatible with freedom, that will be because freedom itself signifies a sense in which an event is not an accident. And again, that will mean that the event depends on or is grounded in something—i.e, has an explanation of some kind. What kind of explanation could that be?

Well, as we have seen, the neutral compatibilist defeats the luck objection by embracing a certain conception of agential control: to control an action, in the sense relevant for freedom, is for the event to have an explanation in terms of the agent's reasons. So it would seem that rationalization, or reasons-explanation, is the relevant form of explanation. As I will argue below, there is good reason to think of reasons-

explanation and intentionality as a form of non-accidentality anyway. But I think this view is also recommended by the fact that it allows us to make sense of the thought that freedom and accidentality are opposed, without needing to appeal to the uniform notion of luck and its metaphorical appeal to luck as an interfering factor which robs an agent of control over an outcome.

Moreover, if freedom is indeed a form of non-accidentality that consists in reasons-explanation, then it follows that an action's being free consists simply in its happening for a reason. Although I cannot definitively argue for that thesis until §5.4, I think that it is an attractive view (and the neutral compatibilist will agree). It allows us to avoid the dead end of claiming that an intentional action may still not be 'up to us' in a sense that cannot be further explained (as proponents of the luck objection often do—see §1.4). On the view I propose, we thus mistake the meaning of the thought that freedom and accidentality are incompatible when we interpret it by saying, as on the uniform conception, that a free action is (1') an intentional action that (2') is not a matter of luck. Instead, being intentional *is* an action's not being a matter of luck in the only sense relevant to freedom. I believe that it is by failing to notice this that the debate on the luck objection has turned into the repetition of moves we have seen in §1.4.

Of course, identifying accidentality with non-intentionality is something we do all the time in ordinary language:

> What is accidental is at least unintended. If I accidentally broke the glass, then I did not break the glass on purpose; in whatever I was doing (unloading the dishwasher or pouring a glass of juice) I did not intend for the glass to break. [Ekstrom 2003, p. 159]

But there is also a deeper reason for identifying intentionality with a form of non-accidentality. Consider Davidson's classic argument in favor of CTA. He points out that an agent may have a reason X for doing φ, and subsequently φ, without X being the *reason for which* she did φ [Davidson 1963, p. 29]. This can occur, for example, when an agent has multiple reasons for φ'ing: for instance, she wants to help her sick neighbor because she wants to do the right thing, while at the same time having a desire for chocolate and a belief that her neighbor will give her chocolate after helping her. Suppose she does help her neighbor. Then the agent can say: 'I didn't help *because* I wanted the chocolate, but because I wanted to do the right thing'—and what she says may of course be true. So it seems that having a reason for φ'ing and actually φ'ing doesn't mean one φ's for that reason.

Moreover, it may even be that one has a reason for φ'ing and one actually φ's, without φ'ing intentionally at all. For imagine that an agent has a reason to raise her arm—for instance, she wants to catch the attention of a friend on the other side of the road. Now if just at that moment her arm would rise due to a queer spasm,

it may be that she succeeds, in that way, in catching her friend's attention. So there is a relation of rational accord between the event of her arm rising, and her mental state—a desire, intention, or other pro-attitude towards signaling her presence to her friend. But *that* this relation of accord obtains is, it seems, merely accidental. It is just a stroke of (good) luck that the agent ends up doing something that she wants to do: that she has a reason to φ is completely accidental to the fact that she actually φ's. Now Davidson correctly saw that in order for a φ'ing to be an intentional action, there must be a non-accidental relation between the actual occurrence of that event and the agent's reasons. That is why he concludes that reasons must be causes of actions [Davidson 1963].[85] An intentional action is an event which is not an accident, because its occurrence is explained by the agent's reasons. And, just as I am suggesting, Davidson himself believed that by giving an account of the non-accidental connection between reasons and action, we also provide an account of freedom [e.g. Davidson 1970, 1973].

So my suggestion is that freedom is a form of non-accidentality: it consists in the dependence of an agent's action on her reasons for acting. This may seem an odd choice of words, because it seems that to say that something is non-accidental is to say that it is *necessary*, i.e., that it happens in all possible worlds. And as a libertarian, I want to deny that free actions are necessitated. But we should note that this understanding of non-accidentality as contrasting to necessity, in *that* sense, is a consequence the uniform conception of luck and the neo-Humean view of modality that underlies it. On the picture I am proposing, we have a variety of forms of explanation, and therefore a variety of modal connections. An action may depend on, or be grounded in, the agent's reasons in a sense that need not imply that it was inevitable that it would occur. That dependence is what the action's freedom consists in. It is harmless to say that freedom is a form of non-accidentality, and thus a form of necessity, if we remember that we do not mean 'necessity' in the sense of something that happens in all possible worlds. This alternative sense is the sense of necessity that, e.g., McDowell has in mind when he expresses the Kantian thought that 'rational necessitation is not just compatible with freedom but constitutive of it' [McDowell 1994, p. 5].

There is thus good reason to think of freedom and intentionality as a form of necessity or non-accidentality.[86] However, in the contemporary free will debate, the connection between intentionality, freedom, and accidentality has largely disap-

[85] This does not mean that I endorse Davidson's view that reasons-explanation is just 'ordinary' causal explanation [Davidson 1963, p. 23]. As I explain in §2.2.1, it is true that reasons-explanation must be causal in a broad sense, but not that it must be the same kind of causation that we find, e.g., in inanimate nature. I further argue for the thesis that rationalization is a *sui generis* kind of explanation in §4.

[86] As I will explain in §2.2.1, that is just to say that freedom is a special *form of causality*, as Kant [1999b, p. 446] puts it.

peared from view. There seem to be two main reasons for this. First there is the prevalence of the uniform conception of luck itself: the assumption that there is a phenomenon of 'luck', pure and simple, has made it impossible to see that there may be different forms of non-accidentality, corresponding to different forms of explanation. Second, as we have already seen (§1.5), almost all parties in the free will debate assume that free action is intentional action that conforms to certain extra conditions. For the libertarian, these extra conditions include at least the requirement that the action be undetermined. Proponents of the luck objection, instead, have tried to argue that 'no luck' is one of the extra conditions. These two assumptions are related. It makes sense to think of 'no luck' as an extra condition that an intentional action must satisfy in order to be free only if we think of luck as a uniform phenomenon. On the other hand, the uniform conception of luck itself encourages us to think that luck must be such an extra condition.

There is thus an unhappy interplay between the uniform conception of luck and the idea that free action is intentionality *plus* some extra conditions. One example of this is the tension in Franklin's view that we have seen at the beginning of §2.1. Franklin admits that it is true that luck excludes freedom, *and* that undetermined actions are 'in the strictest sense imaginable' a matter of luck [Franklin 2011a, p. 216]. It is obvious that he wants to say that there is one sense of luck in which it is incompatible with freedom, and quite another (indeterminism) in which it is not. But it seems that he is unable to clearly disambiguate these two senses because he does not notice that the sense in which a free action must not be an accident is precisely that it must be intentional. And of course, Franklin is an event-causal libertarian.[87] Therefore he cannot admit that an event is free in virtue of having a reasons-explanation is—after all, on the theory of action he adheres to (CTA), intentional action is compatible with determinism. This shows how the almost total dominance of CTA, together with lack on reflection on what it is for something to be a matter of luck in a harmful sense, has led to confusion in the free will debate.

I have argued that if we think of luck as a uniform phenomenon which is understandable independently from freedom, we cannot make sense of the intuitive thought that freedom is opposed to accidentality. We *can* make sense of that thought on the assumption that accidentality is the lack of an explanation for the occurrence of an event. On that view, an action is an accident in a sense that renders it unfree only if it is not intentional. If that is right, then it follows that the intuition that freedom and accidentality are opposed does *not* provide any material for a successful luck objection against the possibility of undetermined free action. For, as everyone seems to agree, undetermined intentional action is possible—the causal connection

[87] See §1.4 for more on Franklin's event-causal libertarianism.

between an agent's mental states and her action *could* be indeterministic.

Of course, the theory of action on the basis of which this is affirmed (CTA) also allows that intentional action is physically determined. Nevertheless, I think the reflections in this section are a step in the right direction for the libertarian. For the answer to the question 'how can an undetermined event be more than just a lucky accident?' turns out to be simply: *by being intentional*. This means that the choice is not necessarily between rampant luck and neutral compatibilism: libertarians can evade the luck objection just by developing an alternative theory of intentional action and reasons-explanation. It therefore seems clear to me that the debate between compatibilists and libertarians *should* be about the credentials of CTA. In the following chapters (especially 4-5), I will take up the task of showing that intentional action requires indeterminism. If my argument in this section is correct, then doing so will *ipso facto* be to show that undetermined action is not accidental in a freedom-undermining sense. However, I must still answer the question why we should not rest content with CTA and neutral compatibilism. Although I will not try to definitely refute CTA until §4.3.1, I will offer some reasons for thinking that CTA does not give a correct account of the kind of non-accidentality exhibited by intentional action in the next section (§2.2).

2.2 CTA and luck

In this section, I will argue that there is reason to doubt whether CTA provides a good account of reasons-explanation, and thus of the form of non-accidentality that underlies *free* action. However, before I can do that (in §2.2.2), I will first reflect on the very possibility of an alternative to CTA in §2.2.1.

2.2.1 A uniform conception of causality?

In §2.1.2, I explained that reasons-explanation constitutes a form of non-accidentality. For as we have seen, such explanation requires more than that an agent has a reason to do what she is doing. If there is not also a causal connection between reason and action, it is just an accident that the agent does something for which she has a reason—it would not be true that she acts *on* the reason. I think this argument is convincing. But how, then, can I at the same time have the ambition to develop an alternative to CTA? Have I not already granted that reasons explanation is causal explanation?

The answer to this question is that I follow Davidson only part of the way. I think his argument succeeds in showing that reasons-explanation must be a *form* of causal

explanation, but not that it must be what he calls 'ordinary' causal explanation. To see why, let us briefly look at Davidson's argument for CTA in more detail. Davidson's argument targets those philosophers who insist that rationalization is more a form of *understanding* than explanation. The idea behind such theories, which we may call *non-causal* theories of action, is roughly that the relation between an agent's reasons and her action is *only* logical: her reasons show why it is rational for the agent to do φ (e.g., to raise her arm), but do not explain the occurrence of the movement (i.e., the fact that her arm goes up).[88] It is against such theories that Davidson argues:

> . . . something essential has certainly been left out, for a person can have a reason for an action, and perform the action, and yet this reason not be the reason why he did it. Central to the relation between a reason and an action it explains is the idea that the agent performed the action *because* he had the reason. [Davidson 1963, p. 28].

So far, I am in agreement with Davidson. What this means is that the notion of a form of explanation, in the sense in which I discussed it in §2.1, is a broadly causal notion. An explanation of why an agent φ's intentionally is more than a way for those who are observing her action to interpret what is going on. That was already implicit in the connection I made between forms of explanation and accidentality: for something to not be an accident, I said, is for it to depend on, or be grounded in something—to have a cause which accounts for it. The relevant notion of explanation is thus not merely subjective: when we give a correct rationalization of why an agent does φ (or equally, that she believes p), we are not *just* giving a story that convinces an onlooker that doing φ (or believing that p) makes sense, but we are saying what actually gave rise to the event in question.

However, Davidson's argument goes further. After all, he aims to establish that rationalization is a species of ordinary causal explanation. How does he argue for that claim? Essentially, his argument is simply that there is no alternative:

> If, as Melden claims, causal explanations are 'wholly irrelevant to the understanding we seek' of human actions then we are without an analysis of the 'because' in 'He did it because . . . ', where we go on to name a reason. [Davidson 1963, p. 29]

I think this is either not an argument at all, or it is a circular argument. Given what we have seen above, Davidson is surely right to rebuke Melden for saying that causal explanation is wholly irrelevant to reasons-explanation, *if* we take that claim as saying that an agent's reasons for raising her arm do not have to explain why her arm goes up. But that does not bring Davidson closer to the claim that reasons cause actions in the same way that, say, the movement of one billiard ball causes another to move, *unless* we presuppose from the outset that this is the only form of causality.

[88] I argue in more detail against such non-causal theories in §5.1.

Of course, Davidson is right that we need an analysis of the relevant 'because'. And interestingly, he claims that Aristotle solved that problem by 'introducing the concept of wanting as a causal factor' [Davidson 1963, p. 29]:

> Failing a satisfactory alternative, the best argument for a scheme like Aristotle's is that it alone promises to give an account of the 'mysterious connection' between reasons and actions. [Davidson 1963, p. 30]

This shows clearly that Davidson presupposes that to say that something is a 'causal factor' must be to say that it is a cause in the sense of what he calls ordinary causal explanation, i.e., event causation. But that is precisely the point I am challenging. Yes, reasons are causal factors—they make it no accident that, e.g., an agent's arm goes up. Yet there can be no inference from that thesis to the idea that all causation is 'ordinary' event causation, of the kind that also occurs in causal transactions between billiard balls. For all Davidson has shown, it may be that there is a variety of forms of causal principles. Indeed, that is arguably precisely the view of Aristotle, who famously discerned four fundamental kinds of causal principles.[89] One of these is, of course, final causation, or what is often called teleological explanation. If it is true that teleological explanation is an irreducible kind of causality, then the bare thesis that desire is a causal factor need not imply that it must be an 'ordinary', more or less mechanistic event-cause.

True, we cannot just stipulate that reasons-explanation is a *sui generis*, teleological form of causality. Davidson is right that an analysis of that kind of causality is required. And of course it must be shown that such causality is indeed the kind that is indicated by the 'because' of reasons-explanation. I will take up these tasks in the chapters to come (especially in 3-4).[90] But it would be foul play to say, as Davidson seems to imply, that in lieu of such an account the default assumption should be that reasons-explanation is ordinary causal explanation. That is just to dogmatically assume what Horst [2015, p. 6] calls a uniform notion of causality. As Horst notes, Bishop [1989, p. 63] admits that CTA is 'largely motivated by the worthy desire for a uniform conception of causality'. However, it seems to me that there is nothing especially worthy about that desire, absent an argument against a pluralism about forms of causation.

Thus it seems that Davidson correctly rejects Hume's doctrine that there are no necessary connections (see §2.1.1). But where Hume started by eliminating a variety

[89] See [Falcon 2015] for an overview.

[90] What does it mean to say that there are different *kinds* of causes? Or: in what sense would, e.g., final causation and ordinary mechanistic causation be varieties of one things, namely, causation? That is a difficult question which I can only answer in the course of my inquiry. In chapter 5, I will argue that the answer is that different kinds of causality—inanimate, teleological, and rational—are different kinds of *powers*.

of distinct kinds of non-accidentality,[91] Davidson brings back only one.[92] If we reject the Humean view that everything is accidental anyway, there is no *prima facie* reason to insist on such a uniform conception of causation and explanation. Undoubtedly, the reason why so many philosophers are attracted to the idea there is only one kind of causal principle is that they believe that this is implied by the scientific worldview. But in fact there is no such implication: it is no result of physics, say, that there is only one kind of causal principle.[93] To establish whether we should be monists or pluralists about kinds of causality, there is no substitute for doing philosophy.

2.2.2 Causal deviance

In the previous section, I have argued that the idea that reasons-explanation is a form of non-accidentality, and thus a form of causation, does not imply that there is no room for an alternative CTA. Rather, CTA gives a specific account of the non-accidental connection between an agent's reasons and her intentional action. That account is reductive in that, according to CTA, that connection—the 'because' of reasons-explanation—is just the 'because' of ordinary causal explanation, and not a *sui generis* causal principle. But as I will now argue, there is reason to doubt whether CTA succeeds in giving a correct account of the relevant form of non-accidentality.

According to CTA, an intentional action is a movement that is caused by mental states whose contents rationally accord with that movement. Thus, it would seem that an action is, roughly, intentional if and only if:

> (1″) the agent has mental states which represents the action as in accord with the agent's desires or goals.
> (2″) those mental states caused the action.

However, that is not the case. For the putative definition of intentional action is threatened by counterexamples known as cases of *deviant causal chains*. Consider the following example of deviant causation by Davidson himself:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to

[91] For example, Hume not only eliminates 'ordinary' causality, but also such things as goodness or moral virtue. See Stroud [2011, pp. 90-124] for a case study of (neo-)Humean attempts to eliminate the latter, and how this compares to their treatment of ordinary causality.

[92] Of course, Davidson's account of causation is itself neo-Humean: causation is, roughly, the instantiation of universal regularities. See Davidson [1967]. In §5.2.1 I argue that such accounts of causation may be unfit even for understanding causality in inanimate nature. If that is right, Davidson would even be wrong about what so-called ordinary causal explanation is.

[93] In §5.1.2, I argue that it *would* be true that there is only kind of causality if determinism would be true on the lowest level. The intuition that the uniform conception of causation is implied by the scientific worldview is thus connected to intuition that this worldview must be deterministic.

> 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. [Davidson 1973, p. 79]

In the climber case, the conditions of representation and causation are satisfied, but the agent nevertheless does not act intentionally. Despite his reasons for loosening his hold, the agent's letting go of the rope is just an accident. This has prompted many philosophers to try to come up with necessary conditions in addition to (1″) and (2″), which together would yield a satisfactory definition that defeats cases of deviance. For example, it is often observed that in the climber case, the problem seems to be that the agent's reasons cause the action through an intermediate state of nervousness. So it may seem that we need an additional requirement:

> ? (3″) The mental states are the proximate cause of the action.[94]

The idea is that a state is not a proximate cause if there are intermediate links (such as the climber's state of nervousness) in the causal chain between that state and the action. But of course, it seems that whenever we act intentionally there almost always *are* such intermediate states between one's putative desire to, say, raise one's arm, and the subsequent arm-rising event. So the question becomes which intermediate links are harmless, and which, like the state of nervousness, undermine the intentionality of the action—which links are 'right' and 'wrong', as it is often put. The proximity condition as formulated here obviously won't do, but there is a huge amount of proposals that attempt to define what a non-deviant causal chain is. I will not discuss and argue against those here,[95] but in §4.2.3-4.3.1, I will argue that the attempt to define the right causal chain is futile anyway. And somewhat surprisingly, Davidson agrees with this assessment:

> Several clever philosophers have tried to show how to eliminate the deviant causal chains, but I remain convinced that the concepts of event, cause and intention are inadequate to account for intentional action. [Davidson 2004, p. 106]

Of course, Davidson did not want to abandon CTA. His idea seems to be that, although we cannot come up with necessary and jointly sufficient conditions for intentional action, the metaphysical reduction of rationalization to ordinary causal explanation still goes through. But that seems doubtful. After all, CTA claims that rationalization is a *species* of causal explanation. Given the *prima facie* difference between the explanation of an action in terms of reasons, and the explanation of a billiard ball's moving in terms of the impact of another ball, it seems we should at least be given an account of the specific *differentiae* of reasons-explanation. In the absence of such an account, we should doubt whether reasons-explanation is really

[94] This strategy is employed, for example, by Mele [1992].
[95] For an excellent survey of such proposals and why they fail, see van Miltenburg [2015, pp. 91-130].

a species of ordinary causal explanation (i.e., of the kind manifested by our billiard balls) at all.

Many proponents of CTA thus try to evade the problem of causal deviance in a different way. They propose that an intentional action is an action that rationally accords with the contents of a mental state that causes the action 'in the right way', or 'non-deviantly'. This is usually considered harmless: the phrase 'in the right way' is often added as an unimportant afterthought.[96] However, I think the need to include such a caveat is not at all harmless for CTA. For notice that the inclusion of this caveat comes down to saying that an action is intentional if it is caused by an appropriate mental state, *and it is not an accident that the mental state causes the action*. For that seems to be precisely what goes on in deviant causal chain scenarios: in e.g. the climber case, it is just a matter of luck that the agent's reasons cause an action that she represents as desirable.

To explain the sense in which it is not a matter of luck that the mental state causes the action, the proponent of CTA must thus appeal to a notion of accidentality that is understandable independently from the concept of intentional action. Indeed, we might imagine a proponent of CTA trying to cash out the idea of non-deviant causation exactly in terms of the uniform conception of luck: it is not an accident that S acts on her desire or intention φ if she also does so in a sufficiently wide array of close possible worlds. That may rule out some cases of deviance—arguably, Davidson's climber would try to keep control over the rope, and thus perhaps he would not let go of it in many close worlds. David Horst [2015, p. 13] considers this possibility of accounting for the 'in the right way' clause, and argues against it in detail. His considerations mirror my argument against the uniform conception of luck (§2.1.1). Even though it may be true that agents with an intention to φ will do φ in a wide array of close possible worlds, that is of no help in devising a definition of intentional action, for 'the connection between intentional action and modal responsiveness runs in the other direction: it is *because* an agent is acting intentionally that the relevant modal condition is true' [Horst 2015, p. 15].

The above suggests that CTA is in a predicament analogous to that of, e.g., the anti-luck epistemologist or the proponent of the luck objection we have seen in §2.1. That is: CTA tries to define a form of non-accidentality, but has to resort to including a 'no luck' clause in that definition. This is problematic, because the point of CTA was to show that certain events are no accident *in virtue* of having an explanation of a specific sort—an ordinary causal explanation which mentions the agent's reasons for acting. But now it seems that in order to define this form of explanation, we

[96] Some examples of those who use the 'in the right way' or 'non-deviantly' caveat are O'Connor [2002, p. 88], Clarke [2003, p. 76], and Wilson and Shpall [2012].

must again take recourse to a notion of non-accidentality. And it seem unlikely that this notion can be defined in non-circular terms. For the sense in which it must not be an accident that a mental state causes an action, in order for the causation to be non-deviant, is of course just the sense in which an intentional action is no accident: the reason must cause the action in such a way as to provide a reasons-explanation of it. Therefore I believe that including a non-deviance clause in the definition of intentional action is, once again, akin to listing the cake on its own recipe.

The underlying problem is nicely illustrated by Davidson's admission that what it is for a reason to cause an action 'in the right way' would be for it to cause it 'through a course of practical reasoning' [Davidson 1973, p. 79]. That seems exactly right: the problem in, e.g., the climber case is that the agent does not *conclude* that he should let go of the rope. The mysterious 'because' in 'she did φ because she wanted to ψ' is, it seems, precisely the 'because' of practical reasoning—an agent's conluding that she ought to φ. But then, what is it for an agent to reason practically to the conclusion that she should φ? On CTA, that is just for an appropriate reason-state to cause her action. So to say that a reason must cause an action through a course of practical reasoning does not define what causation 'in the right way' is. The mysterious 'because' of reasons-explanation and practical reasoning, it seems, cannot be demystified by reducing it to ordinary causal explanation.

Of course, I have not yet definitely shown that it is impossible to give an analysis of the non-deviance clause. As I said, I will not do that until §4.3.1. But I think the above should be enough to make us doubt whether the form of non-accidentality exhibited by intentional action is really just that of ordinary causal explanation.

2.3 Conclusion

In §2.2.2, I suggested that there is reason to think that CTA fails as an analysis of the non-accidental connection between reasons and intentional action. If that is right, then—on the assumption which I share with Davidson, that acting intentionally is acting freely—it follows that CTA equally fails as an analysis of free action. And indeed, I think that is what underlies the intuitive formulations of the luck objection that we have encountered. For consider again Mele's 'neural roulette' analogy:[97]

> As soon as any agent . . . judges it best to A, objective probabilities for the various decisions open to the agent are set, and the probability of a decision to A is very high. Larger probabilities get a correspondingly larger segment of a tiny indeterministic neural roulette wheel in the agent's head than do smaller probabilities. A tiny neural

[97] We have first seen this example in §1.3.

> ball bounces along the wheel; its landing in a particular segment is the agent's making the corresponding decision. [Mele 2006, p. 8]

Interestingly, the neutral compatibilist and the event-causal libertarian tell us that this actually might be more or less how free actions *do* come about. And, I have argued, *assuming* CTA, we cannot fault them for this. For if CTA is true, then actions produced through such neural roulette would be non-accidental in the sense relevant to freedom: they would be intentional actions. In the contemporary free will debate, this is a convincing argument—but only because almost all parties in the debate are committed to CTA.

What I have hoped to show in this chapter is that we are free to reason in the other direction. If we find Mele's roulette model an absurd picture of free will, then perhaps this is because the theory of action which says that the outcome of this random process is an intentional action is mistaken. That is: it may be that we have the intuition that such action cannot be free *not* because of the indeterminism in Mele's example, but because there is something wrong with the picture of the mind as a neural machine. For we should ask ourselves: would it be any better if the neural mechanism was deterministic? I believe not. So perhaps the original intuition behind the luck objection—that freedom is opposed to accidentality—does not militate against the possibility of undetermined action, but against the reductive picture of practical reasoning and action-explanation. Of course, I have more work to do in order to show that this reductive picture is mistaken. However, I think I have sufficiently motivated the need to put the question whether we can reduce reasons-explanation to ordinary causal explanation back on the agenda of not just the philosophy of action, but also of the free will debate.

In the coming chapters, I will operate on the hypothesis that the answer is negative: intentionality is a *sui generis* form of non-accidentality, corresponding to an equally irreducible form of explanation—and that is, as I have argued in §2.2.1, a form of causality. We can now begin to see what that form of causality must be: the causal connection between reasons and action is the agent's practical reasoning. I will develop a non-reductive account of practical reasoning and action-explanation in chapter §4. However, before I can fruitfully do so, it is necessary to first address the question what reasoning in general is. For in contemporary philosophy, the assumption that reasoning is nothing more than the operation of mechanistic causes is so common that it is difficult to see how practical reasoning could be any different. As Patricia Kitcher [2011, p. 70] points out, it is commonly assumed that only phenomenal consciousness, i.e. the qualitative feel of sensations, forms a 'hard problem'[98]

[98]The reference is, of course, to Chalmers [1995] and his discussion of the phenomenal quality of conscious sensations.

for reductionist or materialist accounts. By contrast, the abilities to think and reason are considered to be easy targets for reduction: isn't it simple to model reasoning in mechanical terms, and even to implement such a model in a computer?

Like Kitcher, I want to argue that this is mistaken. There is a different form of consciousness, namely self-knowledge, which is intimately tied to cognition, and which makes it impossible to account for reasoning on reductionist grounds. Before moving on to specifically *practical* reasoning, I will first explain this point for 'ordinary', theoretical reasoning in chapter 3. Reasoning, I will argue, is a *sui generis* causal principle that is essentially *self-conscious*. In chapter 4, I will then argue that the same goes for reasoning in its practical variety: practical reasoning is a form of causality characterized by *practical knowledge*. On the basis of this, I will then argue that CTA cannot escape the problem of accidentality which gives rise to deviant causal chains (§4.2.3). In chapter 5, I finally argue that from this theory of action and practical reasoning, it follows that intentional action is undetermined and free. Freedom consists not in being caused by special mental causes, but in a special kind of *spontaneous* causality.

* * *

Chapter 3

Self-knowledge and reasoning

The argument in the previous chapter revealed that different forms of non-accidentality correspond to different forms of explanation. And, I argued, freedom consists in such a form of non-accidentality, namely, the kind inherent in intentional action. Understanding freedom will thus require us to understand the form of explanation that is constitutive of intentional action: action-explanation. And I suggested that event-causal accounts of free will[99] remain prone to problems of accidentality precisely because they try to reduce the explanation of intentional action to ordinary causal explanation. A proper account of freedom thus requires us to take seriously the idea that action-explanation is a *sui generis* form of explanation.

We have also seen (§2.2.2) that an account of action-explanation is intimately bound up with an account of practical reasoning. In explaining an agent's action, one represents her as reasoning practically. For example, explaining an action by saying 'she is turning left because she intends to go to the restaurant', one represents the agent as deriving one action or intention—turning left—from another. So it seems the kind of non-accidentality described by action-explanation *is* the agent's practical reasoning—her deriving the conclusion from the premise. To defend the idea that action-explanation is a *sui generis* form of explanation, I will thus have to develop a non-reductive account of practical reasoning. That is the combined purpose of the present chapter and the next. As I noted at the end of chapter 2, the idea that practical reasoning may resist reduction to event causation may be surprising to some: in contemporary philosophy of mind, it is often supposed that if anything forms a problem for reductionism, it must be sensation. In this chapter, I will therefore first

[99] This also applies to agent-causal accounts, in so far as they depend on CTA for an account of intentionality (see §1.3.2).

argue that the phenomenon of reasoning itself defies reduction, before moving on to specifically practical reasoning in the next chapter.

It will be helpful to say something about how this chapter relates to the next, and to my eventual aim of giving a libertarian account of free will. In the next chapter, I will present an alternative theory of practical reasoning and intentional action: one developed by Elizabeth Anscombe in her *Intention*. As we will see, Anscombe suggests that central to understanding intentional action is *practical knowledge*, an agent's non-observational knowledge of what she is doing. Her idea is that this kind of *self-knowledge* of what one is doing is what the *sui generis* character of action-explanation and practical reasoning consists in: practical reasoning is a form of causality that is essentially *self-conscious*. I will then argue (in chapter 5) that from this, it follows that intentional action is undetermined and free action.

However, before I can present and defend Anscombe's theory, there is an important obstacle we must first overcome. The idea that self-knowledge can help us in formulating a non-reductive account of practical reasoning will seem misguided to many contemporary philosophers. For self-knowledge is often given an equally reductive treatment as intentional action itself: to self-know that one intends/believes that p, for example, is simply to be in a higher-order belief state that one intends/believes that p, of course suitably caused in a justification-providing way.[100] It is difficult to see how that could contain the seeds of a radically different account of intentional action. In this chapter, I will therefore first investigate the phenomenon of self-knowledge, in order to argue against these reductive understandings of it. Because such contemporary accounts of self-knowledge often focus on *theoretical* self-knowledge, i.e., knowledge of what one believes, I will also do so in this chapter, and leave practical self-knowledge for the next chapter.

The main tenet of contemporary accounts of self-knowledge that I will argue against is what I call the Distinct Existences assumption, or (DE): the idea that believing that p and one's self-knowledge that one believes this are two different thoughts, states, or acts of mind. In place of (DE), I will defend the idea that believing that p and self-knowing this are *one* act of mind. Coming to believe that p on the grounds that q *is* to self-ascribe p *as* something one holds true. If this picture of self-knowledge is right, there also emerges a positive account of (theoretical) reasoning

[100] Authors who take such a higher-order view of self-knowledge are, for example, Evans [1982], Setiya [2011], Morgan [2015]. The commitment to this view is often implicit. That it is not explicitly thematized is, I think, an indication of how pervasive the assumption that self-knowledge consists in a propositional attitude which depicts another attitude really is. The same assumption plays an important role in the debate about *phenomenal* consciousness, i.e., consciousness of sensations or perceptions. There the idea of popular higher-order theories is roughly that, e.g., a sensation is conscious if there is a second-order state with contents like 'I have sensation X'. See, e.g., Rosenthal [2006]. Although I will not go into the debate about phenomenal consciousness, I think such higher-order views suffer from difficulties similar to those that I will argue plague accounts of self-knowledge of thoughts.

as an essentially self-conscious form of causality—one that thus cannot be reduced to what Davidson calls ordinary event causation.

I will start (§3.1) by introducing the topic of self-knowledge and the role that the Distinct Existences assumption plays in contemporary theories. As I will explain, that assumption makes it difficult to understand how self-knowledge could be *transparent*, in a sense to be explained. In §§3.2-3.2.1, I will critically examine an attempt to reconcile transparency and (DE) by Setiya [2011]. Setiya's account is of great interest to my purposes in this thesis, because his ultimate objective (like mine) is to give an account of practical self-knowledge—knowledge of what one *intends*, and arguably, of what one *does* intentionally. But in contrast to my purposes in chapter 4, Setiya wants to make room for practical self-knowledge in the context of an orthodox causal picture of action, that is, 'without disputing the picture of intention as a mental state distinct from and causally responsible for its own execution' [Setiya 2011, p. 177].[101] He hopes to defend the possibility of such practical knowledge by reflecting on the nature of self-knowledge in the theoretical domain—departing, of course, from an equally orthodox causal picture of theoretical reasoning and belief. By understanding the difficulties with Setiya's attempt to reconcile transparency with the standard causal model in the case of belief, I aim to show that in understanding intentional action and practical self-knowledge, we must likewise abandon that model.

In §§3.3, I argue that theories like Setiya's cannot accommodate what I will call *knowledge of grounds*: a subject's knowledge not only of *what* she believes, but also of *why* she believes it. As we will see, in representing herself as believing p on the grounds of q, a subject *ipso facto* judges that p *follows* from q (§3.3.1). In §3.3.2, I will argue that accommodating this insight requires us to abandon (DE) and accept a different view of the relation between belief and self-knowledge. Following Boyle [2011], who calls his theory 'reflectivism', I suggest that believing that p and self-knowing this are one act of mind.

In §§3.4-3.4.1 I will trace resistance to the reflectivist thesis to a certain understanding of first-person reference: the idea that 'I' refers as a quasi-demonstrative. Following Anscombe [1975], I argue that we must reject that idea. First-person thoughts are rather acts of *self-predication*. This allows us to better understand the reflectivist thesis: self-knowledge *is* a subject's act of making up her mind (§3.4.2).

[101] This is in contrast to a number of other interpreters of Anscombe's doctrine about practical knowledge, according to whom 'Anscombe shows what is at best optional in the contemporary view that we understand what intention is only by asking what the mental state of intending is and how it could causally contribute to the production of intentional action' [Moran and Stone 2011b, p. 73]. This radical reading of Anscombe corresponds to the theory of intentional action that I will defend in chapter 4. Setiya correctly ascertains that holding on to the orthodox picture and CTA requires him to resist the radical reading. For as I will show later on (§4.3.2), the thought that there is no fundamental contrast between intending and doing (as Setiya [2011, p. 176] describes this alternative doctrine) is a natural consequence of the rejection of (DE) in the practical domain.

Finally, I argue that this means that (theoretical) reasoning must be a *sui generis* form of non-accidentality: reasoning is a kind of causality that is essentially self-conscious, and therefore *spontaneous* in a sense to be explained.

3.1 Transparent self-knowledge

Among the various things about which a human being may form thoughts, it seems, are facts concerning what that very human being is thinking. Just as one may form beliefs about the thoughts of *other* people, for instance '*she* believes that there is an apple tree in the orchard', so one can think: '*I* believe there is an apple tree in the orchard'. But one's thoughts about one's *own* thoughts seem special in that they do not seem to depend on any of the behavioral or contextual clues that are typically required for forming beliefs about what someone else is thinking. For example, one does not need to infer that one believes it will be warm today from the fact that one is wearing shorts. Rather, we can tell what it is that we are thinking *immediately*, or *without observation*. And this capacity to know our own mind non-observationally seems incredibly robust: for example, there is no room for doubt about *who* it is that believes that there is an apple tree outside, as, looking in the mirror, there may be doubt about whether it is really oneself whose clothes are on fire. That is, judgments about what we ourselves are thinking are *immune to error through misidentification*, as many have noted.[102] And in general, persons enjoy a great deal of epistemic authority about the way their minds are made up: if someone claims that he does not believe that there is an apple tree outside, we will normally not dispute whether that is really what he believes, save special reasons to think he is not being sincere, is deluded, etc.

We may call the kind of thinking at issue *self-thinking*. And when such a thought is (non-accidentally) true, it is *self-knowledge*. The object of a self-thought would normally be linguistically expressed using the first-person pronoun, 'I'. But of course not every 'I'-thought is of the non-observational, authoritative kind described: one who looks in the mirror as his clothes are burning may think, "I am on fire". If he is right, then in one sense, his knowledge is self-knowledge: it is knowledge that essentially concerns *himself* in a way that the thought "the man in the mirror is on fire" would not.[103] However, I will restrict the term self-knowledge to the narrower sense of knowledge of the contents of one's own mind, of the non-observational, authoritative kind.

[102] Perhaps most famously Shoemaker [1968].

[103] The use of 'himself' here is, thus, an instance of the indirect reflexive pronoun that Castañeda [1966] styles 'he*'. Also compare Geach [1957]. See §§3.4-3.4.1 for more on this pronoun.

The epistemological peculiarities of the capacity for self-knowledge have long puzzled philosophers of mind and language. Indeed, many find them so peculiar that they have argued that the capacity to say what one thinks cannot be a capacity to *report* or *judge* anything substantive at all. So-called *expressivists*, for example, claim that our 'avowals'[104] have only the *surface* form of statements about the speaker and his belief, and are really only more or less guarded ways of presenting a claim about *P* itself' [Moran 2001, p. 101]. However, as Moran argues, 'expressivism about a class of utterances (e.g., moral discourse, first-person discourse) prevents the questions of knowledge from even arising' [Moran 2001, p. 102]. As it is my ambition to show that reasoning is a form of causality to which the agent's knowledge is essential, I will have to try to defend a genuinely *cognitivist* theory of self-knowledge.

That there is (at least *prima facie*) an important connection between reasoning and self-knowledge becomes clear from what is now regarded as a commonplace observation about the phenomenology of self-knowledge: that we often (or always) arrive at such knowledge by reflecting not on the question 'do I believe that *p*?', but rather by reflecting on the question whether *p* itself holds. That is, self-beliefs (at least normally) conform to what Moran calls the Transparency Condition. Here is Evans's canonical illustration of the phenomenon of transparency:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward–upon the world. If someone asks me "Do you think there is going to be a third world war?", I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" [Evans 1982, p. 225]

So it seems we arrive at self-knowledge of what we believe by reasoning about what is true. That is difficult to understand: how can we gain knowledge of what is seemingly one state of affairs (that one believes that *p*) by reflecting on whether another fact (*p*) obtains? One way to deal with this difficulty would be to abandon cognitivism: if self-knowledge is not genuinely knowledge about a subject's state of mind (as, e.g., the expressivist claims), the paradox disappears. But again, that would only accommodate the phenomenon of transparency by denying that reasoning about the truth of *p* can lead to self-*knowledge* at all. Moreover, such a move is unnecessary—or so I hope to show. The problem of how to reconcile the phenomenon of transparency with a cognitivist account of self-knowledge is a genuine challenge, but not one that is impossible to solve. In the rest of this chapter, I hope to show that this reconciliation

[104] The term 'avowal' is used in a broad sense to mean any sincere statement about one's own mind, covering both cognitive phenomena ('I believe that/intend to *p*') and sensory experiences (e.g. 'I have a headache'). See e.g. Bar-On [2004]. Since my topic is the more narrow domain of self-knowledge of cognitive states, or acts of thought, I will avoid using this broad term except when referring to others who use it.

is possible, and that it teaches us something important about the relation between self-knowledge and reasoning.

Richard Moran has famously tried to reconcile transparency and cognitivism. The following observation constitutes the cornerstone of his explanation of how reflection on the question whether p can lead to knowledge that one believes that p:

> What conforming to transparency comes to is the commitment that beliefs I call my own are beliefs I can endorse as true. [...] If my intention is to report on my belief as such, and I know (how could I fail to know?) that my belief about X is what I hold to be true of X, then my intention will not be thwarted if I make this report by considering what is true of X. [Moran 2001, p. 105]

Moran's suggestion is that it is because one *knows* that what one believes about X is what one holds true about X that we can have transparent self-knowledge. This solution is, I think, on the right track, but it will not do as it stands. Although it may be (trivially) correct that holding true and believing cannot come apart, this does little to solve the problem we find ourselves with. For it does not answer the question how I can justifiedly form a belief with the content 'I believe that there is an apple tree outside' on the basis of reasons—the scent of apple-blossom drifting through the window—that do not bear *at all* on the contents of my mind. Even if I know that what I believe about X is what I hold true about X, that does not yet answer the question what it *is* that I hold true about X. For example: I think 'the scent of apple-blossom is drifting in from outside'. Now, that cannot be sufficient to ground belief in the proposition that one believes that there is an apple tree outside—for all you know, it may be that the scent of apple-blossom has not convinced you of the presence of an apple tree. After all, it may be there is another hypothesis you find more plausible, e.g., that grandma is outside, wearing her old-fashioned apple-blossom perfume.

Of course, on the basis of the proposition that the scent of apple-blossom is drifting in, you may form the belief that there is an apple tree outside; but how can you conclude from *that* that you *believe* there is an apple tree outside? Even if you believe that p, it does not *follow* from the truth of *what* you believe—namely, p—that anyone at all believes it. That is, there is no truth-value link between the propositions p and 'I believe that p'. Now Moran suggests that one may draw the conclusion that one believes that p, when one believes that p, because one knows (and cannot fail to know) that one's beliefs conform to what one holds true. But that requires one to *know* what it *is* that one holds true—and that is just to know what one believes. Paradoxically, in order to arrive at knowledge of my belief that p by reflecting on reasons bearing on the truth of p itself, I would already have to know what I conclude on the basis of those considerations.

Knowledge of the principle that 'my belief about X is what I hold to be true of

X', then, cannot justify an *inference* from considerations in favor of p to the belief that one believes that p. How then could conforming to the transparency condition be rationally mandated? Moran attempts to add to his account of the justification for the transparency condition by considering the interesting connection between self-knowledge and the essentially *active* attitude that rational subjects take towards their own beliefs:

> [O]nly if I can see my own belief as somehow 'up to me' will it make sense for me to answer a question as to what I believe about something by reflecting exclusively on that very thing, the object of my belief. [Moran 2001, pp. 66-67]

And elsewhere, he writes:

> I *would* have a right to assume that my reflection on the reasons [for thinking that p] provided an answer to the question of what my belief [. . .] is, if I could assume that what my belief here is was something determined by the conclusion of my reflection on those reasons. [Moran 2003, p. 405]

I believe that Moran is correct to seek an account of the justification for transparent self-knowledge in the way our reasoning *determines* how our minds are made up. However, by itself the idea that we actively determine what we *do* think by reasoning about what *to* think, still does little to solve the problem we encountered. For note that, even if I can safely assume that the conclusion of my reasoning on the question, whether p, determines my belief about p, then drawing the inference that I believe that p will still require me to identify p *as* the conclusion of my reasoning—that is, as something that I have come to believe. On the assumption that a proper account of how we make up our minds must be an account of how we know our own minds, both will remain elusive until we understand how the transparency condition can hold. That is to say: to develop a sound account of the activity we deploy in determining what to think, that is, of the activity of reasoning, *is* to close the (apparent) epistemic gap between our reasons for thinking that p and the corresponding self-knowledge. In a slogan: self-knowledge and reasoning are one. That will necessarily sound obscure at this point (although I think it is the root of much venerable philosophical thinking), but I hope to elucidate it in the course of this chapter.

My suggestion is that what generates the paradox of transparency and opens up the epistemic gap is just the seemingly innocuous assumption that to self-ascribe a thought is to have a higher-order belief about that thought—that is, to have the belief-attitude to a proposition describing a propositional attitude, e.g. 'I believe/intend/surmise/hope that p'. On this picture, it is possible that one has, for some time, the belief that there is an apple tree in the orchard, while remaining ignorant of this fact until one is prompted to reflect on the contents of one's own thoughts about the matter. As Descartes says:

> Many are themselves ignorant of their beliefs. For since the act of thought by which we believe a thing is different from that by which we know that we believe it, the one often exists without the other. [Descartes 1931, p. 95]

Let's call this the Distinct Existences (DE) assumption.[105] It seems that there is a strong tension between this assumption and the transparency condition. The transparency condition requires that we gain self-knowledge by reflection on the grounds of our beliefs; but since according to (DE), a self-belief is a distinct propositional attitude with a different content than the first-order belief, reflection on the reasons for our first-order beliefs can never, it seems, justify self-belief. We can never move from the *content* of the first-order belief to the proposition *that* one believes it.[106] For cognitivists about self-knowledge who espouse the transparency requirement, there is a genuine difficulty about reconciling it with (DE). I will eventually argue that this challenge is, in fact, insurmountable, and (DE) must thus be dismissed. However, I will first (§3.2) consider Setiya's [2011] attempt to account for cognitive, transparent self-knowledge while upholding (DE).[107] I will then argue (§3.2.1) that Setiya cannot reconcile them in a way that shows how transparently acquired self-beliefs could be rationally justified. Later, I will explain that this is not a peculiarity of Setiya's account: as we will see, any account which embraces (DE) will have to suffer from similar difficulties.

3.2 The rule of transparency

Setiya's starting point is a rejection of the idea that we arrive at self-knowledge by an *inference* from evidence about ourselves, such as our own behavior or 'quasi-perceptual appearances of belief' [Setiya 2011, p. 178]. That is a promising start in light of the problem we encountered above: it is impossible to infer that we believe that p from reflection on considerations supporting p itself, even if we know that our beliefs are sensitive to those considerations. Instead, Setiya insists that '[k]nowledge

[105] Moran [2001, p. 12] refers to this quote by Descartes, but he does not make clear whether he himself endorses (DE). However, I think it is fair to say that in his [2001], he does not explicitly notice the need to reject (DE), as I will argue we must do if we want to account for transparent, cognitive self-knowledge.

[106] It is worth noting that (DE) only has this result when combined with the transparency condition: on an inner perception model of self-knowledge, for example, we can form the second-order attitude on the basis of a process or mechanism that is independent of considerations supporting the first-order attitude. But of course, such proposals suffer from their own difficulties. For one thing, such theories are arguably committed to the idea that the first person pronoun refers in the manner of a name or a demonstrative, which I criticize in §3.4. Moreover, it is difficult to see how self-knowledge, on the inner perceptual model, could make a difference to theories of theoretical and practical reasoning, or even why it should be an important philosophical topic at all.

[107] As I noted in the introduction to this chapter, Setiya's account is a good target for critique, because his ultimate aim is to make room for practical self-knowledge within a version of CTA. In the next chapter (§4.2), I will argue that Setiya and CTA cannot in fact accommodate practical self-knowledge.

of one's own beliefs is often *groundless*' [Setiya 2011, p. 178]. Of course, we are prone to ask: if the relevant knowledge is not arrived at by inference, how do we arrive at it, and how can it count as *knowledge*? As Setiya says:

> This might be thought enough to generate a puzzle: How is groundless self-knowledge so much as possible? But the force of this challenge is unclear. [. . .] What principle threatens the possibility of groundless self-knowledge? It would beg the question to assume that knowledge always rests on inference [. . .]. Of course, we can ask in general terms when a belief is justified and when it counts as knowledge. But that these questions can be raised is hardly evidence of some skeptical problem for groundless knowledge of belief. A better question—one that points in the direction of transparency—is how it can be *rational* to form beliefs about one's own beliefs not on the basis of perception or inference. [Setiya 2011, p. 179]

That, then, is the task Setiya sets out to accomplish: to explain in what way it can be rational to form self-beliefs that are groundless, in the sense of not depending on observation or inference. That self-belief is groundless, however, does not mean it has no relation at all to the capacity for inference. For, Setiya argues (rightly, to my mind) that is not an open question whether a creature that possesses the ability to form beliefs through rational reflection on other beliefs also possesses the capacity for self-knowledge. That would be the case on, e.g., a quasi-perceptual account of self-knowledge: if knowing our own minds requires an *additional* sensory capacity, then it seems our possession of such a capacity is entirely contingent, relative to the capacity to form first-order beliefs. According to Setiya, there is rather a conceptual connection between having the concept of belief and having the ability to form self-beliefs. Setiya thus claims that 'the impossibility of self-blindness is not a doctrine to be argued for, but a datum in the study of self-knowledge to be taken for granted and explained' [Setiya 2011, p. 180].[108] He thus accepts the following thesis:

> (**Cognitive Self-Knowledge**): If *A* has the capacity for inference and can ascribe beliefs to others, she has the capacity for groundless knowledge of her own beliefs. [Setiya 2011, p. 180]

Given this constraint, there are seemingly just two ways in which a capacity for groundless self-belief might be shown to be rational. First, the capacity for inference might *presuppose* the capacity for self-belief. Or vice versa, the capacity for self-knowledge might be a *consequence* of the capacity for inference. Setiya argues that the first cannot be true: 'in its simplest form, inference is wholly world-directed, moving from premise to conclusion without self-ascription'. [Setiya 2011, p. 182]

[108] I agree with Setiya on this point: again, if our capacity for self-knowledge were a contingently realized mental module of some kind, it would be hard to explain the importance of self-knowledge to a broader account of rationality at all. I argue for the claim that it *does* have this importance in §§3.3—3.3.1.

This must be the case, he argues, because if it were necessary to form a belief that one believes that p in order to infer anything from p, a regress would threaten, and inference could never get off the ground. Thus, although the capacity for inference is, for Setiya, the *source* of the capacity for self-knowledge, it is important to see that exercises of the two capacities remain entirely distinct (in accordance with (DE)).

In what way can the capacity of inference be the source of self-knowledge, while the acquisition of self-knowledge is itself not inferential, but groundless? Setiya argues that the solution to this puzzle requires us to take two steps:

> The first is to recognize inference as a species of *epistemic rule-following*: the application to evidential rules of a more general capacity to form beliefs on the basis of other beliefs. [...] The second step is to form a *rule of transparency for belief* inspired by Evans: "whenever you are in a position to assert that p, you are *ipso facto* in a position to assert 'I believe that p'."[109] [Setiya 2011, p. 184]

This rule of transparency, then, is Setiya's proposed way of accounting for Moran's Transparency Condition. How should we understand this rule of transparency? Setiya warns against a misleading interpretation of the rule, according to which 'I first ask what I am in a position to assert, then make an inference from my answer' [Setiya 2011, p. 184]. For as we have already seen, such an inference from 'I am in a position to assert that q' to 'I believe that q' would be impossible (and any attempt to supplement the reasoning with additional premises such as 'I believe everything I am entitled to derive from premises I hold true' seems hopeless). And anyway, Setiya is after a non-inferential account of self-knowledge.[110] So a better way of interpreting Evans's rule must be sought:

> On a more attractive view, I draw directly on the state in virtue of which I am in a position to assert that p. If I am capable of inference, I have the capacity to form beliefs on the basis of my beliefs. Groundless self-knowledge exploits this capacity, not to form the belief that p-or-q on the basis of my belief that q, or the belief that p in light of evidence that p, but to form the belief that I believe that p on the basis of my belief that p. [Setiya 2011, p. 184]

There are doubts to be had about the epistemic credentials of the rule of transparency, since it seems to differ so much from other kinds of epistemic rules. But let us bracket those concerns for the moment (until §3.2.1). Setiya claims that he has formulated

[109] The quote is from [Evans 1982, pp. 225-226].

[110] Alex Byrne [2011] defends a theory of self-knowledge that is substantially the same as Setiya's. However, ironically, he dubs his theory 'inferentialist', saying that the step from belief to self-knowledge deserves the name 'inference' because it is a way of producing a new belief based on an old one, in a way that is epistemically warranted. As Boyle [2011, p. 230] says, the problem with saying that such a step is inferential is that the inference would be mad. Setiya, wisely reserving the term 'inference' for logically valid derivations, is thus forced to postulate a different form of 'epistemic rule-following'. Of course, the question is whether the step is any less mad when we refrain from calling it an inference. In §3.2.1, I argue that it is not.

an account of groundless self-knowledge that renders the constraint of (Cognitive Self-Knowledge) true. For the capacity to form self-beliefs by following the rule of transparency is 'a repurposing of the capacity to follow rules of inference' [Setiya 2011, p. 184]. So apart from that latter capacity, Setiya argues, 'no further *general* capacity is required' [Setiya 2011, p. 185] in order to be capable of groundless self-knowledge.[111]

However, we may wonder whether there is not a further prerequisite of Setiya's procedure that he overlooks. That is the ability to deploy the first-person pronoun, or the capacity to think 'I'-thoughts. After all, a subject must be able to form a belief with the content '*I* believe that *p*'. This requirement is made explicit in Evans's remarks, made in the context of knowledge of our own perceptual states:

> A subject can gain knowledge of his internal informational states in a very simple way: by re-using precisely those skills of conceptualization that he uses to make judgements about the world. Here is how he can do it. He goes through exactly the same procedure as he would go through if he were trying to make a judgement about how it is at this place now [. . .] The result will necessarily be closely correlated with the content of the informational state which he is in at that time. Now he may prefix this result with the operator 'It seems to me as though . . . ' [Evans 1982, pp. 227-228]

Setiya endorses the procedure that Evans describes. Presumably, then, when applying the rule of transparency for belief, we are exhorted to 'prefix the result with the operator "I believe . . . "'. It is thus a necessary condition of Evans's and Setiya's accounts that we are already able to apply that prefix independently of having the ability to know one's own mind. One must *first* have the ability to think 'I' thoughts *before* one can apply it in the context of Evans's procedure. So it is a consequence of Setiya's account that the capacity for first-person thought and the capacity for self-knowledge are independent of one another.[112]

Note that the possibility of separating the two capacities in this way depends on the assumption (explicitly defended by Evans) that we should conceive of the 'I' in 'I believe that *p*' as analogous to a proper name or demonstrative *X* in '*X* believes that *p*' (an assumption I criticize, based on Anscombe's [1975] argument, in §3.4). That is, it must be assumed that to ascribe a belief to *oneself* is just a special instance of the more general ability to ascribe beliefs to *someone*. The possibility of thinking 'I', or of thinking of oneself, on this view, is no different than the possibility of thinking about another person or object. I will later argue (§3.4.1) that it is this unhappy

[111] The emphasis on 'general' is to allow for the possibility of error in specific situations, in which a subject with the general capacity for inference may yet find herself incapable of following the rule of transparency due to some kind of psychological obstacle.

[112] For more on the distinctness of the two capacities, and its problematic consequences for developing a sound account of self-knowledge, see p. 111.

separation of the topics of first-person thought and self-knowledge that generates much confusion. But for now, we can see why Setiya thinks his account satisfies the constraint of (Cognitive Self-Knowledge): anyone who has the ability to ascribe belief to *someone* thereby has the ability to ascribe belief to *herself*, since the latter is just a special case of the former.

3.2.1 The (ir)rationality of the rule of transparency

As noted, Setiya's goal in describing his 'rule of transparency' was to show how it could be rational to form groundless beliefs about one's own thoughts. So assuming that it is even possible for a subject to form beliefs according to Setiya's procedure, we should ask why forming them would be *justified* in such a way that these self-beliefs could amount to knowledge. The problem is, of course, that in applying the rule of transparency, the first-order belief that p does not justify the resultant self-belief in the way that a premise supports a conclusion—indeed, the first-order belief is *not* a premise at all, since applying the rule of transparency is not supposed to be a matter of drawing an inference. So in what *other* way could forming the self-belief be rational? Setiya retorts:

> As before [. . .] we should ask for the argument behind the question. What makes this appear irrational? Is it that one's conclusion is not based on evidence? Since our topic is groundless self-knowledge, that much is inevitable, and it would beg the question to doubt its legitimacy. [Setiya 2011, p. 186]

This attempt to shift the burden of proof is, I think, unconvincing. It was Setiya's aim to give an account of self-knowledge that would show how it could be rational to form self-beliefs without grounds. He has described a procedure or mechanism by which one might operate on the contents of certain beliefs and form beliefs with the content 'I believe that p', and has suggested that we accept the rule of transparency as a principle that rationally legitimizes this procedure. But of course it is part of Setiya's burden of proof to show *why* we should accept that rule as a fundamental epistemic principle. The objection is not that applying the rule of transparency is irrational because it is not a species of inference from evidence, but rather that we have not been given *any* positive account of why applying it *would* be rational. It is not enough to say that without the rule of transparency, groundless self-knowledge would be impossible—for all Setiya has shown, it may yet be that groundless self-knowledge *is* impossible. So an account of the epistemic credentials of the rule of transparency and the associated procedure for forming self-beliefs is required.

Setiya is unimpressed by this demand for rational justification, and thinks that there is no special problem about 'the non-accidental reliability of beliefs acquired by

transparent means'. Surprisingly, Setiya here interprets the demand for justification as a demand that the rule of transparency be non-accidentally *reliable*. But at the same time claims that 'in saying this, we do not fall into reliabilism. It is essential that self-knowledge of belief be acquired by the exercise of a rational capacity, not just a causal process.' [Setiya 2011, p. 186] So Setiya rejects the thought that all that is required for the exercise of a procedure to be justified is that it reliably yields true beliefs.[113] The procedure must, in addition, be a *rational* capacity in some further sense. However, it is telling that he does not explain in what sense the application of the rule of transparency *is* the exercise of a rational capacity, as opposed to 'just a causal process'. As I will shortly argue, this is because there is no way of explaining that: despite Setiya's denial, his account of self-knowledge *is* an inherently reliabilist one.

But first, it will be instructive to reflect on the reasons why Setiya claims that 'no epistemic rule is more reliable than the rule of transparency.' [Setiya 2011, p. 186] Why should we suppose this? Apparently, it seems, for the reasons Evans himself adduces when describing the procedure for arriving at knowledge of how things perceptually seem:

> This [i.e., forming the belief that it perceptually seems that p by transparent means] is a way of producing in himself, and giving expression to, a cognitive state whose content is *systematically* dependent upon the content of the informational state, and the systematic dependence is a basis for him to claim knowledge of the informational state. [Evans 1982, pp. 227-228]

Transposing this to the case of self-belief: when one concludes that one believes that p by prefixing the operator 'I believe that . . . ' to the result of first-order deliberation about p, the self-belief counts as knowledge because there is a systematic dependence between the two belief-states. Since the procedure is to prefix the contents of a belief one actually has, the resulting self-belief must be true. Evans optimistically concludes (and Setiya agrees):

> If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states: even the most determined skeptic cannot find here a gap in which to insert his knife. [Evans 1982, p. 225]

Now it is interesting to note that the basis for calling the resultant self-belief 'non-accidentally reliable', on the Evans-Setiya picture, resides purely in the fact that in

[113] Such reliabilism, in epistemology, is roughly the idea that knowledge does not require the subject to be able to give a justification of her belief, as long as that belief was formed according to a mechanism which reliably tracks the truth. This reliability is then typically spelled out in terms of safety, or the absence of luck, in the way I discussed in §2.1.1. Reliabilism is thus opposed to broadly internalist theories of knowledge, which demand that the subject be able to know *that* her belief is justified. See [Goldman and Beddor 2015], especially §2 thereof.

prefixing 'I believe that . . . ' to the contents of a belief one *actually* has, one cannot arrive at a false self-belief. For that makes one wonder why it should be true that, in forming self-beliefs, our eyes are 'directed outward upon the world'—and thus whether the rule is aptly described as a rule of *transparency*. The idea behind the transparency condition was that in order to know our own thoughts, we must reflect on the very same considerations that we would reflect on when trying to answer the first-order question whether p. Setiya and Evans present their accounts as ways of accommodating that insight. But now it seems that reflection on the reasons for or against p is not necessary at all: given that one already believes that p, one can operate on the contents of that belief, prefixing them with the 'I believe'-operator, and arrive at a reliably formed true belief in that way. One's reasons for thinking that p, surprisingly, are not relevant for arriving at self-knowledge *at all*. Thus it seems Setiya's rule of transparency is rather inaptly named: in what sense are our eyes directed 'outward' in acquiring self-knowledge according to his procedure?

Thus the 'non-accidental reliability' of self-beliefs, on the Evans-Setiya picture, is not actually linked to the considerations that support one's first-order reasoning. And this raises a further worry about the sense in which these beliefs, arrived at through applications of the rule of transparency, could be rational. For even if the procedure yields reliably true beliefs, this reliability is completely independent of any *justification* that the subject could give. This, it seems, makes for a fundamental difference between the proposed rule of transparency and any other 'species of epistemic rule-following'. Take for example a simple rule of inference like 'when one is in a position to assert p and $p \rightarrow q$, then one is *ipso facto* in a position to assert that q'. When one forms the belief that q by application of this epistemic rule, then one normally *knows* that one is in the position to apply it: that is, one knows that the premises are true, and that this licenses the conclusion. That is of course not to say that the rule of inference must be an additional premise—that would be the mistake for which Lewis Carroll [Carroll 1895] famously warned us. But it is to say that, when forming beliefs through a rational capacity, we are normally able to say *why* we believe what we believe—i.e., to say why that belief is *justified*.

We will examine the importance of the ability to explain our own beliefs in more detail below (§3.3). For now, note the contrast with the rule of transparency. Setiya's formulation of that rule allows us to arrive at the belief that 'I believe that p' from the belief that 'p'. When one applies that rule, one cannot yet know oneself to be in the position described by its antecedent—for that would once again require you to already know how your mind is made up. The rule of transparency must always be applied, as it were, behind the subject's back. So the subject finds herself having beliefs of the form 'I believe that p', but cannot say why it is that she believes this.

If she is philosophically inclined, and has formed the complex hypothesis that self-beliefs are formed via Setiya's procedure, then perhaps she may guess (and hope) that the self-beliefs she finds herself with were produced through that procedure. But there is no way of knowing whether she was ever in the appropriate condition—that of believing that *p*—for the rule to kick into action.[114]

Setiya will perhaps want to reply that the reason why the subject cannot tell why she has the self-beliefs that she does have is that such belief is, after all, groundless. Would it not beg the question to suppose that all rational belief formation requires that the subject has the kind of knowledge that allows her to justify the belief? But it does not seem obvious that if a belief is groundless, in the sense of not depending on observation or inference, it must therefore be a mystery to the subject. It is one thing to say that self-knowledge does not rely on evidence, and quite another to say, as Setiya must, that it is impossible *in principle* for a subject to know whether her self-belief is justified. Moreover, if that were the case, then it would be hard to see why self-knowledge should be expected to play an important role in an account of theoretical and practical reasoning. For the motivation behind the transparency condition was the idea that we are not mere idle bystanders with regards to our own thoughts, and that our activity in reasoning is the ground for claiming knowledge of them. Paradoxically, on Setiya's account, our self-beliefs would be the ones most alien to us—the ones furthest removed from our active capacity to make up our minds.

I hope that it has become clear that, despite Setiya's denial, his account of self-knowledge is thoroughly reliabilist. The only sense in which self-belief is justified, on his account, is a sense that is necessarily untransparent to the subject. Just as typical reliabilist observers would be unable to distinguish perceptual beliefs that are generated by reliable mechanisms from perceptual beliefs rigged by a clever neuroscientist, so a thinking subject would be unable to distinguish her self-beliefs arrived at via the rule of transparency from randomly generated bits of mental content. And as I will argue in §3.3, whatever the prospects of a reliabilist epistemology in the field of perception, there is no place for it when it comes to our ability to know our own thoughts.

[114] Incidentally, this has the peculiar consequence of making it impossible to tell whether Setiya's theory of how we acquire groundless self-beliefs is actually true of any of our self-beliefs. There is a serious question of how we can know any theory of self-knowledge to be correct, if that theory insists that the formation of self-beliefs happens behinds the subject's back. For in that case, the *account* of self-knowledge of course cannot itself arise from self-consciousness. I think the account of self-knowledge I expound in this chapter does better on that score. Although I will not be able to go into detail, I believe that account at least opens the door to saying that we know what self-knowledge is *by having* such knowledge.

3.3 Knowledge of belief and knowledge of grounds

It is a familiar fact that epistemic and practical subjects can not only say what it is that they believe or intend, but can also give the grounds they have for the beliefs and intentions they thus know themselves to have. In the practical domain, this capacity to answer the question 'Why?' concerning one's intentions or actions has been made famous by Elizabeth Anscombe [Anscombe 1963].[115] And as Anscombe observes, the ability to answer the question 'Why do you intend to do *X*?' seems to share the general characteristics of the ability to answer the question 'What is it that you intend?': the relevant knowledge is non-observational, and the subject enjoys a certain first-person authority over the answer. It seems that something similar holds for the theoretical case: apart from knowing *what* they think, believers typically enjoy knowledge of their reasons for thinking it. The purpose of this section is to argue that such knowledge of grounds, as we will call the ability to answer the question 'Why?' in both the theoretical and practical domain, is incompatible with reliabilism about self-knowledge. Reliabilism is therefore a false doctrine.

Those with reliabilist inclinations will perhaps already want to object against the idea that believers *typically* enjoy knowledge of grounds: that might seem like a naive internalist fantasy. In order to sidestep discussions of this nature, I will grant the reliabilist, for the sake of the argument, that there *may* be some beliefs for which we are unable to give grounds. My argument will rather proceed from the thought that we sometimes *do* have knowledge of the grounds of our (first-order) beliefs, and will attempt to show that reliabilism fails to accommodate even this modest claim.[116]

Why should we think that knowledge of grounds forms a problem for reliabilism? We observed (§3.2.1) that application of the rule of transparency would necessarily be untransparent to the subject: a subject can never be aware *that* she is in a position to apply the rule, for that would already require her to know what she believes. It followed that, for any particular self-belief, the subject cannot know that it was actually formed according to the procedure that is supposed to rationally justify it. That is, a reliabilist subject cannot, merely by exercising her capacity for self-knowledge, answer the question 'Why?' about her self-beliefs.

Now the reliabilist may wish to retort that this is exactly as it should be. Her

[115] I will give a detailed account of Anscombe's theory in chapter 4, in which I will also explore the analogy between knowledge of one's reasons for action and knowledge of the reasons for one's belief at length.

[116] In the rest of this chapter, I will thus only consider beliefs which are grounded in inference from other beliefs. This is purely for reasons of scope: of course, there are other ways of arriving at belief, such as, perception and testimony. I therefore do not espouse a version of coherentism—the doctrine that the only thing that can justify a belief is another belief [e.g. Davidson 1986]. However, my argument in the following sections does commit me to the idea that thinkers also enjoy knowledge of grounds of beliefs arrived at in these other ways.

claim, after all, is that self-knowledge is literally groundless. In the case of self-belief, there *are no grounds to be known*. So perhaps self-beliefs are a special case: they do not admit of an answer to the question 'Why?', simply because there is no such answer. But, we might imagine the reliabilist arguing, that needn't mean that subjects cannot answer the question 'Why?' about their first-order beliefs. Imagine a subject saying 'there is no milk in the fridge'. We might ask her: 'Why is there no milk in the fridge', to which she might reply 'because you didn't buy any'. Does it not seem that we have here an example of a subject who is perfectly able to provide reasons for her first-order belief, i.e., possesses knowledge of grounds, without needing to reflect on the origins of any of her self-beliefs?

But to draw this contrast between ordinary beliefs, about which we at least sometimes possess knowledge of grounds, and self-beliefs, of which such knowledge is impossible, would be unfortunate. The contrast implies the following: we are sometimes able to answer the question 'Why p?', but not the question 'Why do you believe that p?'. However, that is mistaken. An answer to the question 'Why p', in the sense relevant to knowledge of grounds, will have to be answer to the question 'Why do you believe that p?'. Consider again the agent who answers 'because X forgot to buy it' to the question 'Why is there no milk in the fridge?'. The fact that X forgot to buy the milk may be a perfectly good reason to believe that there is no milk in the fridge. And it may actually be the correct account of how the fridge has come to be devoid of milk. But that by itself is not enough for the subject's answer—'X forgot to buy it'—to count as knowledge of the grounds of her first-order belief, that there is no milk in the fridge. In general, it may be that q is a good reason for someone to believe p, without it being her reason for believing that p. It may be, for example, that she does not in fact believe q at all. Or it may be that, although she also believes q, she has arrived at the belief that p by way of a different belief, namely the belief that r. In that case, responding to the query 'Why p?' by uttering 'q' does not display knowledge of the grounds of one's first-order belief that q. So the subject's answer 'because X forgot to buy it' only displays knowledge of grounds if it is, in fact, the reason why she came to believe that there is no milk in the fridge.

The question 'Why p?', in the sense according to which being able to answer that question is possessing knowledge of grounds, is thus not the brute demand to be presented with just any proposition that, if it were true, would explain the first-order fact that p. Rather, it is a demand for the subject's actual reason for thinking that p. This is parallel to what we have seen in (§2.1.2) for the case of action. As we have seen (§2.1.2), Davidson argued that an agent may have a reason for doing φ that is nevertheless not the reason why she did it. He argued that a thought (in his case, a belief-desire pair) is the reason *for which* one acts only if it *explains* why the action

occurs. As we have seen, he called such an explanation a rationalization. Similarly, knowledge of the grounds of one's belief that *p* must be knowledge of something that explains why one has come to believe it. An onlooker may give the relevant explanation by saying, e.g.: 'she believes that *p because* she believes that *q*'. The ability to answer the question 'Why?' about one's belief that *p* is the ability to give a true rationalization of it. It is knowledge that, for example, 'I believe that *p* because I believe that *q*'. There is thus no such thing as answering the question 'Why *p*?', in the relevant sense of 'Why?', without at the same time answering the question 'Why do I believe that *p*?'. The problem for the reliabilist is that on that account, subjects would be unable to answer the latter question, and thus they would be unable to justify their (first-order) beliefs.

Note that to say that knowledge of grounds is of the form 'I believe that *p* because I believe that *q*' is not to say that what justifies believing *p* is the proposition 'I believe that *q*': strictly speaking, of course, the fact *that* one believes something does not usually rationally count in favor of any other proposition. It is just to say that the uncanny ability we have to answer the question 'Why?' about our beliefs is an ability to become aware of *which* first-order propositions constitute our reasons for thinking what we do, and that this involves being aware of them *as* things we believe. And this, it seems, is precisely what we should expect if we take the phenomenon of transparency seriously. If transparency means knowing that what one holds true (say, *q*) is what one believes, then when I know *why* I hold *q* to be true—say, because of *p*—then I know that *p* is something I equally hold true, and hence something I believe.

Until this point, I have been proceeded merely on the plausible assumption that cognizers *sometimes* possess knowledge of grounds. We can now make this idea more precise. We have seen that to answer the question 'Why?' about a belief is to give a true rationalization of it. It would therefore seem that there is a close connection between the capacity for knowledge of grounds and the capacity of inference. For a true rationalization is of the form 'I believe that *q* because I believe that *p*', which identifies the *reasoning* behind one's belief that *q*. The idea that we sometimes possess knowledge of grounds can thus be explained by saying that, at least sometimes, when one makes a certain inference, one *knows that one has made it*.

Furthermore, it seems that this capacity to know that one has inferred *q* from *p* shares the essential characteristics of self-knowledge: the relevant knowledge is non-observational, and, of course, non-inferential (it is impossible to *infer* that one has made the inference '*p*, so *q*' from the fact that one believes that *p* and *q*). We have seen that Setiya says that 'the impossibility of self-blindness is not a doctrine to be argued for, but a datum in the study of self-knowledge to be taken for granted

and explained' [Setiya 2011, p. 180]. This motivates him to endorse his principle of (Cognitive Self-Knowledge), which states that possessing the capacity for inference is sufficient for possessing the capacity for self-knowledge. Given the similarities between the peculiarities of ordinary self-knowledge and knowledge of grounds, one wonders why the latter should not be given a similar foundational status. And in fact, it has recently been suggested by Matthew Boyle that the fact that we possess non-observational knowledge of grounds is a constraint on any sound theory of reasoning. He formulates a principle he derives from Moran, and hence calls Moran's Constraint:

> (**MC**) My reasoning '*p*, so *q*' must normally put me in a position to know that I believe that *q* because I believe that *p*. [Boyle 2011a, p. 12]

Just as reasoning from a premise *p* to a conclusion *q* sometimes yields knowledge that we believe that *q*, so (MC) states, the very same reasoning sometimes[117] yields knowledge that we believe *q* because of *p*.

It is, I think, fairly obvious that reliabilism cannot accommodate this principle. For how could it be, on the reliabilist account, that making an inference yields knowledge of grounds? As Setiya argues, '[i]n its simplest form, inference is wholly world-directed, moving from premise to conclusion without self-ascription' [Setiya 2011, p. 182]. When one makes an inference, one simply forms the belief that *q* on the basis of the belief that *p*. To come to believe, additionally, that one believes that *q*, an exercise of the wholly distinct ability for self-knowledge will be required. But knowledge of grounds *does* require self-ascription: when one knows that one believes that *q* because one believes that *p*, one thereby knows that one believes that *q* and *p*. So if making an inference is wholly-world directed, it cannot yield knowledge of grounds.

A reliabilist, then, must reject (MC). In order to explain why knowledge of grounds of grounds is *ever* possible, she would thus have to argue that there is some capacity in addition to the capacity for inference, which registers the occurrence of an inference. Thus the capacity for knowledge of grounds and the capacity for inference would be related in the same way as a capacity for self-knowledge, on e.g. an internal perception model, would be related to a capacity for belief: the capacities are logically separate, and there is no reason why creatures who possess the one capacity should also possess the other. As we have seen (§3.2), Setiya rejects internal perception models of self-knowledge because they have this consequence. And it is hard to see why an analogously separate capacity for knowledge of grounds would be any more acceptable. So the reliabilist seems to have no principled grounds to reject

[117] The principle (MC) is stronger: it says the we do not *sometimes*, but *normally* gain such knowledge. That this is indeed the normal case will become clear from the following argument.

(MC)—and yet she must reject it.

Moreover, I argue, there are very strong reasons to accept (MC) even apart from this. To see this, consider a philosopher who denies that we can have knowledge of grounds. We are surely entitled to ask her: 'why do you believe that?'. And paradoxically, any answer to that question will display the very capacity that is being denied existence. That is not just an amusing coincidence, but shows that the ability for knowledge of grounds is very intimately tied to the ability to hold a belief on the basis of reasoning at all. As Boyle [2009, p. 151] points out, without knowledge of grounds, we would be unable to participate in 'the game of giving and asking for reasons'.[118] Without knowledge of grounds, we would not understand how the different beliefs we hold hang together. And arguably, that is just to say that we would not be able to reason at all. For reasoning, it seems, *is* understanding that one thing follows from another. If that is right, then (MC) must be true, and so we must reject reliabilism.

There is another way of bringing this out. Imagine a subject who thinks: 'q is true because p is true'—or equivalently 'p, therefore q'. What she thinks is not just that $p \to q$, nor does she merely view p as good evidence for q. Rather, she thinks both that p is true, and that this settles it that q is true. In thinking 'q because p', she thinks of p and q as connected in a manner that does not leave it open whether q is true. Rather, she will think of q as something she holds true—i.e., as something she believes. Thinking that q is true because p is true implies that one views one's belief that p as the explanation for one's belief that q. So thinking 'p, therefore q' implies that one has knowledge of grounds. And that is just what (MC) asserts. Thus, the possibility of thinking 'p is true because q is true' requires that (MC) is true. Yet the reliabilist must reject (MC). Therefore, it seems, reliabilism destroys the possibility of reasoning.

3.3.1 Knowledge of grounds and reasoning

I have argued that a correct account of knowledge of grounds must conform to (MC), and that a reliabilist account of self-knowledge must be rejected because it cannot do that. In order to make progress towards an adequate alternative account of self-knowledge, we should begin by noting that reliabilism does not do justice to what was the core idea of transparency: that in acquiring self-knowledge, 'one's eyes are, so to speak, or occasionally literally, directed outward–upon the world' [Evans 1982, p. 225]. For as we have already seen, the Evans/Setiya procedure completely severs the connection between one's first-order deliberation and one's self-beliefs. The rule

[118] Boyle attributes this expression to Brandom, who, apart from using it in his own [1994], also attributes it to Sellars in his notes to Sellars' *Empiricism and the Philosophy of Mind* [Sellars 2000, p. 123].

of transparency that Setiya offers does *not* require one to consider one's reasons for believing p, but simply allows the formation of the relevant self-belief whenever the belief that p is present. A more fruitful approach to self-knowledge that could hope to accommodate (MC) would have to restore the connection between transparency, inference, and knowledge of grounds. To restore that connection, it is important to revisit the idea that we can come to know that we believe that p by deliberating on the first-order question whether p.

We have seen that Moran suggested that the possibility of transparent self-knowledge depends on our ability to actively make up our minds. For a subject to believe that there is a relation between her first-order deliberation and self-knowledge, Moran claims, she would have to 'assume that *what* my belief . . . is was something determined by the conclusion of my reflection on those [first-order] reasons'. How could a subject assume that? Boyle's answer is that she could only assume this if she had knowledge of grounds:

> If Moran is right, the sort of agency I exercise when I deliberate must be one that normally puts me in a position to know, on the basis of my drawing the conclusion that Q, that I believe that Q. Moreover, it seems that a related point must apply to my knowledge of my *grounds* for drawing that conclusion: if I reason 'P, so Q', this must normally put me in a position, not merely to know *that* I believe that Q, but to know something about *why* I believe Q, namely, because I believe that P and that P shows that Q. If I could not assume that *all* of these commitments undertaken from the standpoint of deliberation correspond to first-order 'matters of psychological fact,' then I could not assume that I am reasoning from my present view of things to further beliefs which will become parts of this total view. [Boyle 2011a, pp. 11-12]

Boyle explains that for a subject to know that she has made up her mind that q is to know that q is something she has arrived at by reasoning from other things she holds true. So it seems that the ability for transparent self-knowledge that one believes that q requires the ability for knowledge of grounds: it requires that one is able to identify one's reasoning *as* one's reasoning. But then it starts to look as if reasoning is not particularly helpful in explaining the possibility of self-knowledge. What we wanted to understand was how one can identify one's beliefs as one's beliefs. Moran suggested that the answer was: because one's beliefs are the product of one's first-order reasoning. Yet now it turns out that, in order for that to explain the possibility of self-knowledge of what we believe, we would *already* need to be able to say what our reasoning is. This seems paradoxical.

I want to argue that the paradox arises because we assume that, as Setiya [2011, p. 182] put it, inference, in its simplest form, is wholly world-directed, moving from premise to conclusion without self-ascription. If that were true, then there

would always have to be a cognitive step involved in moving from the reasoning to knowledge that one has reasoned—and we do not understand how such a step could be justified any more than a step from 'p' to 'I believe that p'. To resolve the paradox, we have to rethink the relation between the capacities for inference, knowledge of grounds, and self-knowledge. What our findings suggest is precisely that *in* reasoning from p to q, we are *already* conscious of doing so. So it is wrong to say that it is one thing to make an inference, and another to self-ascribe it. Reasoning 'p, so q' and coming to know that one believes that q because one believes that p must be *one* act of a single cognitive capacity. In what follows, I will try to explain what this means.

A good place to start is an objection of Boyle's against the idea that inference could be a reliable capacity, the operations of which could in principle be untransparent to us:

> . . . a belief, once formed, doesn't just sit there like a stone. What I believe is what I hold true, and to hold something true is to be in a sustained condition of finding persuasive a certain view about what is the case. Even if we grant that a disposition to pass from one content to another could deposit various arbitrary beliefs in my mind, those beliefs would be unsustainable if I, understanding their contents, could see no reasonable basis for holding them true. [Boyle 2011b, p. 231]

Boyle suggests that it is essential to belief that we possess knowledge of grounds *throughout* the span of time during which we hold it. Our ability to say *why* we believe that p is a necessary condition of our being able to continue to hold p true—the belief could not be sustained without it.[119] So if Boyle is correct, continually holding p true—an actualization of the capacity to form beliefs through inference—*cannot exist* without knowing, at the same time, that one believes that q because one believes that p. The point can be put by saying that it is a *logical* feature of belief that it dies out without the support of reason. The fact that we tend to lose conviction in propositions for which we cannot give any grounds is more than an empirical regularity.[120] Without the ability to say why one believes that p (or if one prefers: to the extent that one is unable to say so), one loses the ability to hold true that p. Hence Boyle concludes:

> [. . .] A (personal-level) inference is not a mere transition from a stimulus to a response; it is a transition of whose terms I am cognizant, and whose occurrence depends on my—

[119] Of course, our belief that p may also persist when we no longer believe the original premise, as long as we then have new reasons for believing it.

[120] That is, to the the extent that it *is* a regularity. Perhaps it is possible to induce subjects to believe all kinds of things for which they have no grounds, through various kinds of psychological trickery and indoctrination. However, it is clear that in such cases, we would precisely be dealing with irrational beliefs. The point is that if and when a subject gives up belief in the conclusion of an inference because he gives up belief in the premise, that is no accident.

> in the normal case: persistently—taking there to be an intelligible relation between these terms. This is what makes it possible for inference to leave me with a sustainable belief: I can see (what looks to me to be) a reason for it. [Boyle 2011b, p. 231]

It indeed seems to follow that inference does not deposit beliefs in our mind that sit there like a stone. A belief that p is sustained by the knowledge that one (still) believes p because one (still) believes that q, i.e., knowledge that one still subscribes to that inference. So an inference is not *over* when one first starts to believe that q. If one infers p from q, one's recognition that q is a ground for p remains at work for the duration of the belief that p. Inference is a capacity which causes a subject to acquire a new belief (the conclusion), and continually sustains it.[121] The causality of the inference—its work in sustaining belief in the conclusion—does not die out until the concluding belief does. This suggests that the causality of the inference does not work behind the subject's back. What sustains the subject's belief is her *recognition* that she believes p because she believes q. So the causality of the inference works *through* the subject's knowledge of grounds.

The capacity for inference is thus best thought of not as a capacity to *form* a belief that p (and then let it sit there like a stone), but as a capacity for (continually) *believing p for reasons q, r, ...*. If that is right, it seems we have a partial explanation of how reasoning can deliver knowledge of grounds. In making an inference, we continually represent q as a premise from which it follows that p. That representation is what the causality of the inference consists in. We need to better understand this special causality of inference. To begin that task, let us first consider an objection.

The objection is that it is mistaken to think that the subject needs to be aware of the way the premises of an inference continually sustain belief in the conclusion. The objection grants that one's continually believing that p is causally sustained by whatever beliefs one inferred it from. Moreover, it grants that this fact puts one in a position to answer the question 'Why do you believe that p?' for as long as one holds p true. But it denies the idea that it is one's knowledge of grounds—one's *awareness* that the premises support the conclusion—that sustains belief in the conclusion. So our capacity for knowledge of grounds may remain doormat, and does not need to be exercised in order for the inference to do its work. Knowledge of grounds thus relates to the capacity of inference in the same way as Setiya thinks the capacity for self-knowledge relates to it: possessing the capacity for inference is a sufficient condition for possessing the capacity for knowledge of grounds. Yet there is nothing

[121] I am speaking of inference as a 'cause' in the broad sense of causality I explained in §2.2.1—it is a form of dependence between two (or more) of the subject's beliefs. Inference is a form of non-accidentality. That inference is such a (broadly) causal notion was already implicit in the idea that knowledge of grounds is knowledge of the explanation of why an agent has come to believe that p. When someone, e.g., believes that p because she believes that q, the 'because' signals that the belief that p is grounded in the belief that q. This belief-explanation represents the agent's reasoning: 'q, therefore p'.

that says that the latter capacity *must* be exercised whenever the former is—we need not *have* self-knowledge in order to have a belief (in line with the (DE) assumption).

If this objection is correct, then we cannot understand the capacity for knowledge of grounds by understanding the causality of an inference, for the latter will not, as i hypothesized, be essentially conscious. Rather, the causal work of the premises in sustaining the conclusion will be one thing, and consciousness of that causal work will be another. And this is indeed how the relation between premises and conclusion of an inference is usually thought of in contemporary philosophy. With Davidson, it is assumed that this relation is one of 'ordinary event causation'. Inferring q from p is for (the onset of) one's belief that p to cause the (onset of) one's belief that q. When we explain someone's belief that q by saying 'she believes that q because she believes that p', we represent that causal relation between those events. First-personal knowledge of grounds will simply be the ability to give an explanation of that form, with 'she' replaced by 'I'.[122]

However, separating the causality of an inference and a subject's consciousness of that causality in this way ruins Boyle's insight that inference does not deposit beliefs in one's mind like a stone. On the standard picture, we are dealing with two events: the premise-belief and the conclusion-belief. It is, on this picture, true that the premise-belief is the cause of the *entire* conclusion-belief, for however long it may last. But it is true to say this only in the same way that, if the assassination of Franz Ferdinand is the cause of World War I, it is the cause of something that lasted (happened to last) from 1914 to 1918. The causal work of the premise-event is *over* once the conclusion-event starts. Just as it may be true to say that the Battle of the Somme happened because the archduke *was assassinated*, so it will be true to say that a subject believes that q because she *believed* that p.[123] But that does not capture the sense in which a subject's grounds for a belief underlie it throughout its duration. The relation between premises and conclusion must put the subject in a position to answer the question 'Why?', and as Boyle says,

> The relevant why?-question does not inquire into the explanation of his *coming*, at some past time, to hold the belief in question, except insofar as the subject's knowledge of how he came to hold the belief speaks to the reasonableness of his continuing to hold it now. Our interest is not in his psychological history, but in the present basis of his

[122] Thus, on the Davidsonian account, belief-explanation is not essentially first-personal in the sense I argue for in §3.4.2.

[123] The present argument is due to Boyle [2011a], who defends the idea that believing must be an active exercise of the capacity to reason at length. Note that the problem for the event-causalist would not disappear if it is stipulated that the premise remains causally active for an extended period of time (perhaps like a table continually supports a vase sitting on top of it). What needs explaining is *why* the belief that p needs this continual support of the premise. If a belief that p where just a mental state or event, as the causal theory conceives of it, there would be no reason for that—it would be an accident that belief in the conclusion dies out when one loses belief in the premise.

> conviction.[124] Nor do we merely expect a person to be able to speak for the reasons why he *shall* henceforth believe *P*; we expect him to be able to speak to the question why he presently *does* believe it, and we hold him accountable for the reasonableness of his answer. [Boyle 2011a, p. 11]

The question 'Why?' asks for a reason that shows why one *presently* holds a certain belief. That is necessary if knowledge of grounds is connected to belief in the way we discussed, so that the ability to say why one believes p is constitutive of being able to view p as something one holds true. To explain why one presently believes that p, it is not enough to say that, at some earlier point, one *believed* that q. For one may have changed one's mind about q, and if so, q cannot be one's ground for presently believing p (once again: a rationalization of p shows more than that p follows from some other consideration—it shows that p is true because q is true).

But, the objector might counter, can we not stipulate that a rationalization 'she believes that p because she believes that q' is only true if 1) her belief that q caused her belief that p and 2) she still believes that q? This proposal does not improve matters. The fact that she still believes that q does not seem to alter the *relation* between premise and conclusion, which is the same whether or not the subject still believes the premise. That is, the causality exerted by the belief that q on the belief that p is the same, whether or not a q-belief continues to be present after the onset of the belief that p. Any duration of the belief that q is *accidental* to the causal connection described. And that makes it impossible to see how an exercise of the relevant causality, i.e., inferring p from q, could yield knowledge that one *still* believes the premise. Even knowledge that one *has* exercised it would leave open what one presently believes.

Knowledge of grounds would not be possible if the causality of an inference would not, by itself, make it no accident that the conclusion of an inference is believed as long as the latter is believed.[125] This continuous sustaining of the conclusion by the premise *just is* the former's being inferred from the latter.

How can it be that this continual work of the premise in sustaining the conclusion yields knowledge of grounds? Recall the result of §3.3. There we saw that by rejecting (MC), we would lose the ability to think 'p, therefore q'. And as I suggested, that means we would be unable to reason at all: reasoning, I said, *is* understanding the connection between premises and conclusion. Without representing q as following from p, there would be no inference, and hence there would not be the peculiar causality we are now considering. The causality obtains if and only if the subject

[124] As Anscombe would say in the context of intentional action: the sense of the question 'Why?' is not that of asking for a 'mental cause'. See Anscombe [1963, §10] and my discussion in §4.1.

[125] This is not to say that one might *never* lose belief in the premise while retaining belief in the conclusion. For example, one might learn that the original premise is false, and at the same time learn that something else is true that also supports the original conclusion, so that one retains belief in the latter.

represents it. This suggests that making an inference *just is* an act of *representing* the premise *as* the ground of the conclusion, or equivalently, representing the conclusion *as inferred* from the premise.[126] An inference, then, would be an essentially self-conscious form of dependence between two beliefs: in making an inference, one represents oneself as making it, and it is *through* that representation that belief in the conclusion is sustained.[127] It would then be no accident that inference yields knowledge of grounds. Rather, having knowledge of grounds and inferring will be one act of a single cognitive capacity. That this is indeed how things stand I will finally argue in §3.4.2. In §3.3.2, I will first make more clear what conception of the relation between self-knowledge and belief underlies this proposal.

3.3.2 Reflectivism about belief and self-knowledge

I have suggested that the capacity for inference is essentially self-conscious: in reasoning '*p*, so *q*', we not only acquire the belief that *q*, but we are also conscious of making the inference. This may seem puzzling, for it can seem as if inference leads us to acquire, in addition to the conclusion, a second belief with the content that the inference has been made. And I started by saying that there can be no valid inference from *p* to 'I believe that *p*': there always remains the additional step from the truth of *p* to the fact that *p* is believed by a certain person, namely, 'I'. The puzzle can be dissolved by seeing that my answer to the question how knowledge of grounds is possible—that making an inference involves a recognition that one is making it—is *not* to be understood as an account of the additional step from *p* to the self-belief that *p*, but as an attempt to show that such an extra step is not necessary. That is what it means to say that drawing a conclusion in inference and knowing that one does this is *one act* of a single capacity: a further step is not necessary. So it is misleading to say that inference must result in two distinct beliefs.

What leads us to say that inference must result in two distinct beliefs—the conclusion, and a belief representing the inference—is that we are used to thinking about the relation between belief and self-knowledge along the lines of the (DE) assumption: belief and self-belief are two distinct acts of thought, e.g. a first-order and a second-order propositional attitude. If self-knowledge is a form of second-order

[126] Kitcher [2011b] gives a succinct presentation of Kant's argument for the same thesis, that an act of inference just is one's knowledge of making it. She argues that judgment would be impossible if we would not be conscious of the grounds for our judgment *in* making it: if we could not know that what we think is based on reasons, we could never come to think it at all.

[127] In §2.1.1, I suggested that knowledge is a form of non-accidentality. For a belief to count as knowledge, I said, is for it to have a certain kind of explanation—an explanation which entitles her to believe it. The account of inference I am describing here is an example of such an explanation. For in making an inference, a subject arrives at a belief that *p* in a way that allows her to say why it is right to believe (provided the inference is not faulty). It will thus arguably be no accident, in the sense required for knowledge, that the subject believes that *p*.

belief, a propositional attitude distinct from the first-order attitude which it is about, then for inference to result in self-knowledge *must* be for it to produce two different belief-states, and the question remains how the second belief—the one representing the inference—can be justified. To understand the way in which inference, in a single act, results in belief in the conclusion *and* knowledge of grounds, we must explicate a different understanding of belief and its relation to self-knowledge.

Again, it will be instructive to take our cue from Boyle's account of self-knowledge, which he calls 'reflectivist'. He explains the difference between a reflectivist account and more orthodox accounts of self-knowledge[128] as follows:

> The reflectivist rejects an explanatory demand that many theorists of self-knowledge accept. He denies that . . . being in a given mental state *M* and believing oneself to be in *M* are two distinct psychological conditions, and consequently denies that the task of a theory of self-knowledge is to explain how these conditions come to stand in a relation that makes the latter knowledge of the former. We can call any account that accepts this conception of the task of a theory of self-knowledge an *epistemic approach*. [Boyle 2011b, p. 235]

An epistemic approach seeks an account of the step from the belief that *p* to the belief that one believes that *p*—Setiya's reliabilist account, with its appeal to the rule of transparency as a reliable epistemic rule, is a prime example. This strategy only makes sense if it is supposed that there is such a step to be made, and thus, that first-order thought and self-knowledge are distinct 'psychological conditions'. Boyle contrasts this with the strategy of the reflectivist:

> Reflectivists [. . .] offer a different sort of account, one that is primarily metaphysical rather than epistemological. [. . .] The reflectivist's task is to explain the nature of various mental states in a way that clarifies why their existence implies that their subject has tacit knowledge of them, and what this tacit knowledge can amount to. [Boyle 2011b, p. 235]

Boyle says that to believe that *p* is to possess *tacit* knowledge that one believes it. This is different from saying—as any (in Boyle's terms) epistemological account must—that someone who believes *p* merely *potentially* possesses knowledge of that fact. When one believes that *p*, the knowledge that one believes *p* is already there: it merely remains, as it were, unattended to in consciousness. Moreover, the claim is that this tacit knowledge is essential to what it *is* to believe something. One's tacit knowledge of believing *p*, for the reflectivist, is not distinct from the belief it is about. What makes it seem as if they *are* distinct is that we are used to thinking that, if a subject knows that *p*, it follows that she is in a belief-state that *p*. So then, if she

[128] Boyle's primary target is Alex Byrne's 'inferential' account of self-knowledge, which I noted in fn. 110 is substantially the same as Setiya's.

knows that she believes that p, her self-knowledge must be one belief-state, and her belief another. However, what our findings suggest is precisely that there is way of knowing that p which is not being in a belief-state that p, namely, where p is 'I believe that __'. Of course, in one sense it is perfectly innocent to say that someone who self-knows that p believes something about herself. However, to suppose that this must imply that she is, in addition to believing that p, also in a different belief-state would be to beg the question against the reflectivist's account of belief. On that account, self-knowledge of believing p does consist in having a certain belief—just not in a belief distinct from p.

Of course, none of this is to deny that we sometimes become conscious of the fact that we believe something, where previously we were not so conscious. That is just the difference between *tacitly* knowing that p and, for lack of a better word, *occurrently* knowing it: in the same way, someone may know that England has a Queen throughout the day (and even in her sleep), without being conscious of it all the time. Since even first-order beliefs may be tacit, the fact that we do not always have the thought 'I believe that p' before our mind is no argument for the view that self-knowledge must be a second-order belief. Rather, becoming conscious of one's previously tacit self-knowledge is a matter of 'turning our attention from what we are representing to the mode of our own activity in representing it' [Boyle 2011b, p. 227].

The account of belief that follows from our reflections on the connection between knowledge of grounds and inference, then, is that to believe that p *is* to self-know that p. Boyle calls this 'being in a state of knowingly believing that p'. This sounds clear enough, but in an important respect it is misleading: to suggest that one's self-knowledge and one's belief are not distinct is to say that believing that p is not a (static or passive) *state*. After all, beliefs are not deposited in one's mind like a stone. Instead, believing p is a matter of continually *holding* p true. For that suggests more than to say that beliefs normally persist over time. It suggests that the *way* they persist is by our performing a kind of *activity*: that of holding p true for certain reasons. And now the crucial idea is that self-knowledge and (first-order) belief relate in the following way: the activity a subject exercises in believing that p *is* her *knowing* what she thinks.[129] They are two sides of the same coin, or, to put things in more traditional philosophical vocabulary: one's belief that p is the *content* of which one's self-knowledge is the *form*.

On the reflectivist picture, it is thus no longer necessary to suppose that in order for inference to lead to knowledge of grounds, it must result in two beliefs: the conclusion, and a belief that one believes the conclusion on the basis of the premises.

[129] I further develop this idea below in §3.4.2.

For according to reflectivism, one's self-knowledge of believing p is not an act of mind separate from one's belief that p. Rather, a subject's believing that p is 'an enduring actualization of her capacity to hold a proposition true for a reason she deems adequate' [Boyle 2011a, p. 22]. And holding p true for reasons q is nothing other than representing oneself as believing p because one believes that q—it is having knowledge of grounds. Believing that p *is* representing p as part of one's total view of reality, relating in some rational way to other things one believes.[130]

The present suggestion will seem alien to philosophers wedded to the contemporary orthodoxy that views beliefs as mental states, and reasoning as the obtaining of an ordinary event-causal relation between such states. However, I argue, this just shows that the contemporary orthodoxy goes wrong in uncritically assuming that we can have a grasp on the concept of belief (or of thought in general) without understanding self-knowledge. For the orthodoxy depends on the idea that beliefs *are* deposited in one's mind like a stone: coming to believe that p is comparable to a mental switch being flipped, after which one is in the state of believing p until, perhaps, the switch is flipped back. That is arguably an overly simplistic view, for it ignores the fact that beliefs do not float about independently. One's beliefs are part of a *totality* of thoughts: believing that p is judging that p fits into one's total view of the world. And what, on the orthodox view, could make it the case that the numerous switches flipped in a certain location form such a totality? That is, what makes it the case that these states form the total view of *one* thinker? Or again, in virtue of what are two such states beliefs of one and the same thinker—of one 'I', rather than of two, or more?

This question—the question of the unity of the thinker—is one that has disappeared from contemporary philosophy of mind.[131] And to those committed to the orthodoxy, it may not even seem obvious that there is a question to be asked here at all. If the states are states of, say, one and the same brain, is it not obvious whose beliefs we are dealing with? However, the question is not whether a given state belongs to this thinker *or* that thinker, out of a given range of individuated thinkers, but what makes *one* thinker in the first place. Therefore it is *not* obvious that all the putative belief states in one physical space belong to a single 'I'. Indeed, why might there not be an 'I' corresponding to every subset of the putative belief-states in this brain, or this room?

It seems that a correct account of the unity formed by a subject's thoughts will

[130] Note that the capacity of inference need not be the *only* capacity to hold a proposition true for reasons the subject deems adequate: perhaps there are others, such as perception and testimony. Compare fn. 116.

[131] An important exception is Kitcher [2011a], who gives a Kantian account of the unity of the thinker, and connects it to issues in modern philosophy. As the below will show, I am broadly sympathetic to her outlook on this question.

have to appreciate the fact that this unity must be one that the subject herself can understand. This is, I believe, what underlies the famous Kantian dictum:

> It must be possible for the 'I think' to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me.[132] [Kant 1999c, B131-B132]

But can this insight not be accommodated on the orthodox picture? Setiya [2011, p. 187] claims that his account does so: after all, each belief can be the target of a second-order belief. So each belief can potentially be represented as part of the subject's total view of reality. However, this will not do. Setiya's account, or any other account that adopts (DE) for that matter, does *not* do justice to the Kantian thesis. For on that account, representing myself as having a certain belief is an act distinct from the belief itself. And then there is nothing that settles it that the second-order thought belongs to the same totality of beliefs as the first-order thought. Thus, there is nothing that settles it that it is the *same* 'I think' that can accompany each representation. And the true point of Kant's thesis is that this will not do—it cannot be an open question to which thinker a certain thought belongs. Thus, in thinking a thought, a subject must *already* be conscious that is *her* thought. If that consciousness—her self-knowledge—would have to take the form of a second-order belief, it would come too late.[133] Instead of illuminating Kant's thesis, Setiya's account unjustly assumes that the concept of (first-order) belief is *more* clear than that of self-knowledge, and thereby obscures the true relation between the 'I think' and one's representations: a belief is only a belief in so far as I represent it as part of my total view of the world.[134]

Now Setiya might object: is it not obvious that each second-order thought belongs to the same thinker's total view? For the second-order thoughts contain a referring expression, 'I'. If we want to know to which thinker's totality of beliefs a particular second-order thought belongs, we have simply to determine the referent of that expression. Two second-order thoughts belong to the same totality of beliefs if and only if the referent of 'I', in both thoughts, is the same person. This line of defense may seem simple and obvious, but in fact it commits a fundamental error. For it pretends that 'I' refers in the way of a demonstrative or a proper name. In the coming sections (§§3.4-3.4.1), I will examine this account of the first-person pronoun and argue that

[132] I am quoting the translation which Setiya himself [2011, p. 187] uses, by N. Kemp Smith.

[133] Compare O'Brien [2007, p. 24], who attributes the same point to Kant.

[134] This implies that self-knowledge is *constitutive* of the thought it represents. Compare §3.4.2. As Kitcher [2011b, p. 66] notes, the thinker's consciousness of the unity of her thoughts must thus be a consciousness which *constitutes* that unity: 'In engaging in cognition, the understanding also partially creates a rational subject'.

it is mistaken. This will open the way to a more in-depth understanding of the reflectivist thesis (§3.4.2), and with that, of what is wrong with the orthodox view of belief and reasoning.

3.4 Self-knowledge and first-person reference

As we have seen above (p. 82), Setiya's account of self-knowledge depended on an important assumption about the capacity to deploy the first-person pronoun, 'I'. To summarize: applying the rule of transparency required us to prefix the contents of a belief-state with 'I believe that'. The capacity to ascribe beliefs to someone—what Setiya calls 'having the concept of belief'—is, on the reliabilist account, a necessary precondition of the capacity for self-knowledge. So that capacity must be understandable independently of the capacity of self-knowledge. First we have the ability to think thoughts of the form '*X* believes that *p*', of which '*I* believe that *p*' is an instance. Given that ability, the reliabilist can go on to explain how beliefs of the form 'I believe that *p*' may be reliably true, and thus amount to knowledge. The ability to ascribe a belief to oneself must thus be thought of as an instance of the more general ability to ascribe a belief to *someone*. In the schema '*X* believes that *p*', 'I' is just one of the objects that can be filled in for *X*. That is, 'I' in 'I believe that *p*' plays the same role as 'Dr. Lauben' in 'Dr. Lauben believes that *p*'—the logical role that Frege calls *Objekt*. The reliabilist about self-knowledge (and as we will see below, anyone whose adopts the (DE) assumption) is thus wedded to the view that the first-person pronoun can be understood as an expression analogous (or indeed identical) to a proper name or demonstrative—that is, an expression that makes reference to an object, namely, the speaker or thinker.

This analysis of 'I' belongs to a long tradition, stemming from Frege himself.[135] However, it faces some important difficulties, notably, how to account for the *sense* of 'I'. Frege taught us that referring expressions have both a meaning (*Bedeutung*) and a sense (*Sinn*)—the former, roughly, corresponding to the extension of the term, and the latter being the 'mode of presentation', or the way through which reference to the object is made.[136] And of course, one and the same object might be referred to using different modes of presentations. So a speaker might use two sentences, 'Dr. Lauben is *F*' and 'Gustav is *F*', to refer to what is in fact the same object, without realizing that Dr. Lauben and Gustav are the same person. But it is difficult to give an account of the sense of 'I'. Suppose that both Gustav and Gottlob think 'I am *F*'. The 'I' in Gustav's thought refers to Gustav, and the 'I' in Gottlob's thought refers

[135] Notably, from his *Der Gedanke* [1918]. For the English translation, see Frege [1956]).
[136] The *locus classicus* for the sense and reference distinction is, of course, Frege [1892].

to Gottlob. This leads Frege to postulate that 'everyone is presented to himself in a 'particular and primitive way' [Frege 1956, p. 298] in which he is presented to no one else—that is, 'I' has a different sense for each thinker. As is well known this leads to difficult problems about the representation and communicability of first-person thoughts. For if Gottlob were to report on what Gustav is thinking, under what mode of presentation could Gottlob think of Gustav? Say that Gottlob thinks: 'Dr. Lauben thinks that Dr. Lauben is *F*'. This may be false, because it may be that Gustav does not know that he is Dr. Lauben (for example, when Gustav is not aware that he has just been awarded his doctorate). What Gustav thinks is neither 'Gustav is *F*' nor 'Dr. Lauben is *F*', but precisely 'I am *F*'. But Gottlob cannot represent Gustav's thought by thinking 'Gustav thinks that I am *F*', since 'I', when used by Gottlob, refers to Gottlob, and not to Gustav. This seems to undermine an important tenet of any broadly Fregean theory of thought: that thoughts are objective, in the sense that one and the same thought can be grasped by different thinkers. Moreover, it makes it impossible to see how we could explain others' intentional actions, which often (or, if we are correct, always) have a first-person component, e.g. 'He is running because he thinks that *he himself* is late for school'.

However, if one rejects the idea that 'everyone is presented to himself in a special and primitive way', we are arguably pushed in the direction of the opposite problem. Suppose the sense of 'I' is one that can in principle be grasped by any speaker. For example, we might suppose that the sense of 'I' is equivalent to something like 'the person who is using this expression'. Then, it seems, a speaker might fail to realize that it is *her* who is uttering the sentence in question. But it seems undeniable that a competent user of the first-person pronoun *cannot* be mistaken about who is meant by 'I'—as it is often said, first-person utterances are 'immune to error through misidentification' [e.g. Shoemaker 1968]. So where Frege's solution makes the sense of 'I' too specific by tying it to a single speaker, a quasi-demonstrative approach suffers from a lack of specificity—it leaves open too much to settle it that the speaker cannot be ignorant or mistaken about who it is that is indicated by 'I'.

Of course, there may still be various more or less successful ways to solve the problems associated with either strategy. Detailing and refuting the various complex ways philosophers have tried to defend them would require a treatise of its own. However, for my purposes it is interesting to see that the fundamental assumption leading to these difficulties is that 'I' is to be understood as a referring expression. As I will argue, rejecting that assumption will allow us to better understand how a non-reliabilist account of self-knowledge is possible (along the lines of §3.3.2), and to see why it is necessary. One philosopher who famously rejected the idea that 'I' is a referring expression is G.E.M. Anscombe [1975]. Let us briefly review her argument.

Anscombe departs from the recognition that what we need from an account of the first-person pronoun is an account of its *sense*. It may be *true* that 'I' is 'the word which each of us uses to speak of himself', but this specification of its meaning is 'hardly an explanation', since the 'himself' in that expression must itself be explained in terms of 'I' [Anscombe 1975, p. 22]. To see why, we must distinguish between the ordinary reflexive pronoun 'his/herself', and the so-called indirect reflexive pronoun. As an example of the former, consider 'with "Dr. Lauben", Gustav refers to himself'. Above, we have seen that this may be true even when Gustav does not realize that he is Dr. Lauben. The ordinary reflexive pronoun leaves open whether, thinking 'Dr. Lauben is *F*', Gustav also thinks '*I* am *F*'. Therefore, 'the word each of us uses to speak of himself' does not account for the peculiar role of the first-person pronoun, when 'himself' is the ordinary reflexive pronoun. If, on the other hand, we interpret the 'himself' here as the indirect pronoun—the pronoun that Geach [1957] and Castaeda [1966] famously identified, the latter formalizing it using the expression 'he*/she*' or 'he himself/she herself'—it is true, but vacuous, to say that 'I' is the word each uses to refer to himself. For this pronoun is simply the indirect-speech analogue of 'I' in direct speech. That is, 'Gustav thinks that *he himself* is *F*' represents the state of affairs that Gustav thinks: 'I am *F*'. Therefore, Anscombe says:

> We seem to need a sense to be specified for the quasi-name 'I'. To repeat the Frege-point: we haven't got this sense just by being told which object a man will be speaking of, whether he knows it or not, when he says 'I'. Of course that phrase 'whether he knows it or not' seems highly absurd. His use of 'I' surely guarantees that he does know it! But we have a right to ask *what* he knows; if 'I' expresses a way its object is reached by him, what Frege called an 'Art des Gegebenseins', we want to know what that way is and how it comes about that the only object reached in that way by anyone is identical with himself. [Anscombe 1975, p. 23]

Anscombe's strategy is to show that the antecedent—'if 'I' expresses a way its object is reached by him, what Frege called an 'Art des Gegebenseins"—is false, because there is no way in which we can explain that 'the only object reached in that way by anyone is identical with himself'. Her argument is by elimination: if 'I' were a referring expression—an expression through the use of which one 'reaches' an object—it would have to function either as a name, or a demonstrative. Anscombe argues that 'I' functions differently from both. We will now see how she does so.

To show that 'I' cannot be a proper name, Anscombe devises an example of a group of speakers. Each speaker has a public name, *B-Z*, worn on their torsos. When seeing another member acting in a certain way, members of the community will report this by using the name they observe on the moving body. For instance, when *B* sees *C* moving, *B* will say '*C* is doing such-and-such'. But they each also have

a name stamped on the inside of their arm. The name happens to be the same for everyone: '*A*'. When a member of the group makes reports about her own actions, 'which one gives straight off from observation' [Anscombe 1975, p. 24], she uses the name she sees on her arm. For instance, one might see an arm marked with *A* moving, and report '*A*'s arm is in such-and-such a position'. Each member therefore uses '*A*' to refer to herself, just as the name 'I' is putatively a name for everyone, which everyone uses only when talking about herself.

In such a case, Anscombe suggests, one of the speakers may see a neighbor's arm, mistake it for her own [Anscombe 1975, p. 24], and say '*A*'s arm is in such-and-such a position'.[137] The consequences to be drawn from this thought experiment remain implicit in Anscombe's text. It seems clear that one problem with the proposed reduction of 'I' to '*A*' is that, unlike 'I', '*A*' allows for the possibility of referring to an object other than oneself. However, Anscombe hints that there is a more fundamental problem:

> In my story we have a specification of a sign as a name, the same for everyone, but used by each only to speak of himself. How does it compare with 'I'?—The first thing to note is that our description does not include self-consciousness on the part of the people who use '*A*' as I have described it. [. . .] This—that they have not self-consciousness—may [. . .] seem not to be true. *B* is conscious of, that is to say he observes, some of *B*'s activities, that is to say his own. He uses the name '*A*', as does everyone else, to refer to himself. So he is conscious of himself. [. . .] But when we speak of self-consciousness we don't mean that. [Anscombe 1975, pp. 24-25]

Anscombe says that not every knowledge of oneself is self-knowledge. When Gustav knows that Dr. Lauben is *F*, he knows something about an object that is, as it happens, identical with the knowing subject. But this identity is not part of what he knows. Or rather, if he does know it, that will be the content of a further thought establishing the identity—'I am Dr. Lauben'. So the problem is not just that one may refer to the wrong object with '*A*'. Even if the speaker *does* refer to the right object, i.e. if the arm she sees really is the speaker's arm, then the speaker only *happens* to refer to herself. Her knowledge is not self-knowledge because the way she refers to herself—her use of the name '*A*'—leaves open whether the object she refers to is *herself*.

If 'I' is not a proper name *is* a referring expression, then, Anscombe argues, it must be a demonstrative. Now demonstrative reference depends on a perceptual relationship to the object: the 'mode of presentation' of a demonstrative is the object's being given to the speaker in perception.[138] But that cannot be the way in which 'I'

[137] Of course, by imagining that such a speaker may *mistake* another's arm *for her own*, we are supposing that the members of our imagined community possess self-consciousness. That is, they must be able to think 'I'-thoughts in addition to 'A'-thoughts.

[138] Anscombe's notion of a demonstrative thus seems to exclude reference to objects by saying, e.g., 'there

refers, Anscombe argues, because demonstratives are prone to reference-failure:

> [. . .] there may be reference failure for 'this', in that one may mean 'this parcel of ashes' when there are no ashes [the parcel is empty, DO]. But 'I' [. . .] is secure against reference failure. Just thinking 'I . . . ' guarantees not only the existence but the presence of its referent. It guarantees the existence *because* it guarantees the presence, which is presence to consciousness. But note that here 'presence to consciousness' means physical or real presence, not just that one is thinking of the thing. [Anscombe 1975, p. 28]

If 'I' is a demonstrative, it must be a strange one indeed: it must be such as to guarantee that the thing one refers to is always present to consciousness. This soon leads philosophers to postulate obscure objects—'selves'—which, being immaterial, are always before one's mind. And thus, Anscombe argues, the idea that 'I' is a referring expression leads to the postulation of a Cartesian ego:

> Thus we discover that *if* 'I' is a referring expression, then Descartes was right about what the referent was. [. . .] Our questions were a combined *reductio ad absurdum* of the idea of 'I' as a word whose role is to 'make a singular reference'. I mean the questions how one is guaranteed to get the object right, whether one may safely assume no unnoticed substitution, whether one could refer to oneself 'in absence', and so on. The suggestion of getting the object right collapses into absurdity when we work it out and try to describe how getting hold of the wrong object may be excluded. [Anscombe 1975, p. 31]

Anscombe's argument is thus an attempt to show that the idea that 'I' refers leads to the absurd idea of a Cartesian ego. However, her argument is often met with scepticism. The consensus is that she has failed to establish that there is no possible way in which 'I' could refer.[139] For example, might 'I' not function more like an indexical such as 'here', rather than like a demonstrative such as 'this' [e.g. Garrett 1998, p. 103]? Or might there not be a form of inner bodily perception, immune to reference failure but still fully material? As I will explain in §3.4.1, these objections underestimate the power of Anscombe's argument. But it is also important to note that what seems to drive these arguments is mostly an incredulity about Anscombe's conclusion, that 'I' does not refer. Where does this incredulity stem from? As Sebastian Rödl points out,[140] it is often said that there is an important connection between reference and *predication*:

is a yellow car outside' (where one hasn't seen or been told about the car). Whether or not we should classify such cases as demonstrative reference does not seem to matter: Anscombe's argument against the idea that 'I' is a demonstrative—that such reference does not guarantee that there is a referent—still applies.

[139] For example, see Evans [1982], Cassam [1997], O'Brien [2007].

[140] My understanding of Anscombe's thesis, outlined in this section and the next, is deeply indebted to Rödl's account. See especially Rödl [2007, pp. 123-126].

> Reference is commonly represented as an act that contrasts with and complements an act of predication: in order to predicate a concept of an object, it is necessary, in an act distinct from the predication, to single out from a manifold of objects the one that must satisfy the concept if the thought is to be true. [Rödl 2007, p. 125]

In order to judge that an apple is red, one must first pick out that apple from the manifold of objects around one. Otherwise, there is no way to determine whether one's statement is true or not. On this conception, what it means to refer to something is to pick out a particular thing as the object of one's thought. This may explain why many philosophers are so skeptical of Anscombe's conclusion. For if picking out an object is a precondition of being able to predicate, i.e., say something about, an object, then it seems to follow from Anscombe's thesis that first-person thoughts and statements do not predicate anything. So if we can point out a way for Anscombe to deny that 'I' refers, while maintaining that first person thoughts are genuinely predicative, we will succeed in quelling an important source of scepticism about her argument. I will do so in the next section.

3.4.1 Receptive reference and subjectless predication

As I will argue in this section, Anscombe indeed does not mean to deny that first person thoughts predicate. What she does deny is that in first person thought, the subject needs to *reach out* in order to pick out the right object to predicate something of. Such an act of reaching out may fail to find a referent at all (as it does for some definite descriptions, e.g. 'the present King of France'), or it may fail to get hold of the *right* object—for instance, when the 'A'-user mistakes someone else's arm, stamped with 'A', for her own. And as Anscombe argues, these mistakes are obviously not possible in the case of 'I':

> It seems clear that if 'I' is a 'referring expression' at all, it has both kinds of guaranteed reference. The object an 'I'-user means by it must exist so long as he is using 'I', nor can he take the wrong object to be the object he means by 'I'. (The bishop may take the lady's knee for his, but could he take the lady herself to be himself?) [Anscombe 1975, p. 30]

Anscombe's argument is that 'I' thoughts are immune to reference failure and misidentification, and that this implies that 'I' cannot refer. It is important to see that she thus has a very specific concept of reference in mind, namely, that of *receptive* reference. As Rödl explains:

> Anscombe's claim that there is no first person reference deploys a concept of reference according to which referring to an object is relating to it in a way that gives application to the notion of getting the object right. This is *receptive reference*, reference mediated

> by an act of receptivity. In the fundamental case, receptive reference depends on a perceptual relationship with the object. [Rödl 2007, p. 124]

Take demonstrative reference, which is a prime example of receptive reference. A subject sees an apple and says: 'This apple is red'. A while later, perhaps after having closed her eyes for a bit, she sees an apple again and says: 'This apple is red.' Both times, the demonstrative 'this' refers through a perceptual relationship to the apple. But the way we refer to the apple—via our perceptual relationship with it—does not guarantee that we refer to the same apple on both occasions.[141] This is obvious when we imagine that, between the first and the second time one looks, the apple has been surreptitiously replaced with another specimen. So as there are two acts of perception, there must be two acts of reference:

> . . . when my receptive link to an object has been broken and reestablished, the nature of that link [. . .] does not fix it that it is the same object from which I receive representations on both occasions. Hence, there are two acts of reference, and the identity of the object is the content of a separate judgement. [Rödl 2007, p. 124]

Anscombe's argument, then, is that perception cannot be the mode of presentation underlying 'I'-thoughts, because it is essential to perception that it gives rise to a notion of 'getting the object right'. This is the fundamental reason why, for her, no quasi-demonstrative account is possible. But is this correct? Many philosophers have objected that, if 'I' is a demonstrative referring through a perceptual relationship with one's own body—being equivalent to something like 'this body here'—no reference failure or misidentification is possible. For (on the assumption that we are embodied beings) whenever there is someone who intends to refer demonstratively by using 'I', there will be a body present that can serve as the referent. And it seems that it will always be the *right* body—for is it not altogether too fanciful to suppose that some *other* body might surreptitiously take the place of the body one means to refer to, as in the case of our two apples? Anscombe counters with another thought experiment:

> [I]magine that I get into a state of 'sensory deprivation'. Sight is cut off, and I am locally anaesthesized everywhere, perhaps floated in a tank of tepid water; I am unable to speak, or to touch any part of my body with any other. Now I tell myself 'I won't let this happen again!' If the object meant by 'I' is this body, this human being, then in

[141] Brian Garrett is one of those who is not convinced by Anscombe's negative argument, that 'I' cannot be referring expression because that would imply the possibility of misidentification or failure of reference. He insists that '. . . there is no reason why we cannot explain the guarantee of sure-fire reference by citing the self-reference rule' [Garrett 1998, p. 102]—that is, the rule that 'I' refers to the speaker. For, Garrett claims, '[i]f Anscombe is implying that the very idea of guaranteed reference is some sort of oxymoron, then argument to that effect is required.' The above reflections on the nature of receptive reference provides the argument that Garrett demands. Of course, Garrett is free to insist that receptive reference is not the only form of reference. But then an account of a form of reference is required that shows why that form is immune to misidentification. As I argue below, such an account is possible, but accepting it comes down to accepting Anscombe's proposal.

> these circumstances it won't be present to my senses; and how else can it be 'present' to me? But have I lost what I mean by 'I'? Is that not present to me? Am I reduced to, as it were, 'referring in absence'? [Anscombe 1975, p. 31]

Again, many philosophers respond with skepticism. Might there not be some form of introspection or proprioception left to provide the necessary reference [e.g. Morgan 2015, p. 1804]? Or is it not enough that the subject has memories of previous sensory experiences [e.g. Sorell 2005, p. 27]? And even if it is right that one cannot refer demonstratively to oneself in the sensory deprivation tank, can we not say that 'I' *normally* does function in that way [Evans 1982, pp. 215-216]? In each case of these cases, 'I' would still refer through a receptive relation to one's body. But once again, this scepticism misses the more fundamental point Anscombe is trying to make. To see this, let us grant that misidentification is out of the question for a demonstrative such as 'this body here'. Still, this does not show that 'I' refers receptively. On the contrary, the reason why we need not fear misidentification of our own body is precisely that we have a way of thinking of ourselves that is *not* mediated by an act of perception.

Suppose that X refers to X via a perceptual relationship with X's body, thinking what amounts to 'the owner of this body here is F'. Now of course, if X makes the same judgment at some later time, she can be very certain that the body to which she is perceptually related at $t + 1$ is the same body as the one to which she was perceptually related at t. She need not fear that the object has been surreptitiously replaced, because the body located in the space she occupies at any t will of course be the same body (namely, *her* body—the body of one and the same person).[142] But what is the basis for her certainty about this? It is, of course, that she *already* knows that it was the same person, *her herself*, who made both judgments. That is what makes the question whether it was the same body that was located *here* (where she is) at t and $t + 1$ look ridiculous.[143] And if the knowledge that it was the same person who made both perceptions is the basis for her confidence that the body to which she is perceptually related now is the same body to which she was perceptually related earlier, then it seems that there must be a way for her to think of herself that is not

[142] That this is not the kind of guarantee we are looking for can, I think, be seen by returning to Anscombe's community of A-users. Suppose that there is, for some unfortunate reason, only one A-user left alive. Then we know that, in fact, when she sees an arm stamped with an 'A', it will be her arm. But this does not mean that the mode of knowledge she employs—perception—is such as to guarantee that she gets hold of the right object.

[143] Of course we do not regularly go about making inferences of the form 'The body in this place I occupy must be the same as the body in the place I occupied a second ago'. But it would have to be that way if perception were the only way we would have of knowing about our bodies. In fact, our cognitive relationship to our bodies is often non-observational. Interestingly, Anscombe's first example of the class of non-observational knowledge (which I will discuss in detail in §4.1) is 'knowledge of the position of my limbs' [Anscombe 1963, p. 13].

mediated by a perceptual relationship to her body. She must be able to think: '*I* saw that this body here was *F* at *t*, and that it was still *F* at $t+1$'. This, I think, is the gist of Anscombe's famous tank argument. In thinking 'I won't let this happen again', one manifests (among other things) the knowledge that one has been in a different state in the past, and that one may (or will) be in a different state later. This is knowledge of a (cross-temporal) unity that cannot be mediated by the senses.

If this is right, then it follows that the Evans/Setiya project of constructing a reliabilist account of self-knowledge was doomed from the start. For if Anscombe's argument is correct, 'I'-thoughts must manifest a way of thinking about oneself that is not mediated by a perceptual relation to one's body. That way of thinking, which allows one to think of oneself even in the absence of any sensory input, is, of course, self-knowledge: non-observational knowledge of (among other things[144]) one's own thoughts. But if self-knowledge is the kind of knowledge, or 'mode of presentation',[145] through which 'I' thoughts relate a subject to herself, then we cannot account for self-knowledge in the way that Evans and Setiya suggest. For as we have seen, their account requires that we take a belief-state and prefix it with the 'I believe that'-operator. So to achieve self-knowledge of what we are thinking, we would *already* have to possess the ability to think 'I'-thoughts: 'I' must refer to the subject in a way that does not *itself* depend on acts of self-knowledge. The fundamental problem with the Evans/Setiya strategy is thus that it tries to develop an account of first-person reference and an account of self-knowledge in isolation from each other.

What Anscombe insists on is thus that in first-person thought no identification of the object of one's thought is necessary. It does not follow that statements of the form 'I am *F*' are not genuinely predicative. Rather, it shows that in first-person thought, there is no need (nor room!) for an act of reference *in addition to* an act of predication. As Anscombe says: 'I'-thoughts are 'unmediated agent-or-patient conceptions of actions, happenings, and states' [Anscombe 1975, p. 36]. That is, they are conceptions[146] of a subject's own thoughts (e.g. 'I believe that p', 'I intend to φ') and feelings (e.g. 'I am cold', 'I am in pain'). It is no accident that the former category, the unmediated agent conceptions, are the objects of the form of knowledge we have

[144] As we will see below, Anscombe claims that self-knowledge includes both 'agent' and 'patient' conceptions of oneself—that is, both knowledge of one's own thoughts, broadly conceived to include acts of both theoretical and practical reasoning, and knowledge of subjective states like pain.

[145] The term 'mode of presentation' is of course a synonym for 'sense', which is misplaced if Anscombe is right that 'I' is not a referring expression. However, as I argue below, there is a broader notion of sense and reference available on which Anscombe could agree that self-knowledge is the mode of presentation underlying first-person reference.

[146] It is difficult to precisely spell out the notion of a 'conception' Anscombe is using here. It is precisely *not* the notion of a thought *about* the subject's thoughts and feelings. The best we can do is to say that such a conception is the subject's unmediated *knowledge* or *consciousness* of her thoughts and feelings—a knowledge or consciousness that is not distinct from those thoughts and feelings themselves. See below (§3.4.2).

been considering in this chapter: self-knowledge.[147] For if one has self-knowledge that *someone*, e.g., believes that *p*, there is no room for the question *who* it is that believes *p*. Self-knowledge is a form of knowledge that does not give rise to a notion of getting the object right.

Now Anscombe says of these unmediated agent-or-patient conceptions:

> These conceptions are subjectless. That is, they do not involve the connection of what is understood by a predicate with a distinctly conceived subject. The (deeply rooted) grammatical illusion of a subject is what generates all the errors which we have been considering. [Anscombe 1975, p. 36]

Anscombe's thought is that 'I', in statements of the form 'I am *F*', does not, after all, play the same logical role as 'Dr. Lauben' in 'Dr. Lauben is *F*.' First-person thought does not consist of *two* elements—a Fregean *Objekt* of which a *Begriff* is predicated. It consists of only one: the *self-predicating* of the concept. To illustrate, consider how we are to depict the state of affairs that obtains when someone knows: 'I believe that *p*'. According to, e.g., Evans, Setiya, and Frege, that state of affairs is to be analyzed as follows:

> D.O. knows that ([he] believes that *p*)

Here, a two-place predicate (knows) is joined with an object (D.O.) and a proposition, which itself joins an object [he] and a predicate, 'believing that *p*'. Of course, the million dollar question is what to substitute for the placeholder [he]. We cannot substitute 'D.O.', nor any other name or demonstrative, for D.O. may fail to know that *he himself* is named by that name or indicated by that demonstrative. As Anscombe says, it is the grammatical illusion of a subject that generates all the difficulties about the purported sense of 'I'.[148] The alternative is to analyze the state of affairs as follows:

> D.O. self-knows believing that *p*

Or, substituting *F* for the predicate 'believing that *p*':

> D.O. self-knows being *F*

Here, 'self-knows' indicates the *way* in which D.O. predicates the concept of being *F*. In acts of first-person thought, the concept of being *F* is not joined to 'a distinctly

[147] I will focus only on the category of unmediated agent conceptions of oneself. It is obvious that the kind of self-knowledge manifested in unmediated patient conceptions (e.g. knowledge of being in pain) is different from the kind of self-knowledge discussed in this chapter, which, I have argued, is closely connected to reasoning. Boyle [2009] argues that the latter variety of self-knowledge, manifested in unmediated *agent* conceptions, is more fundamental than self-knowledge manifested in *patient* conceptions: for creatures like us, the ability to say that we are, e.g., in pain takes the shape of a *rational* ability—the ability to conceptualize such sensations.

[148] And as Anscombe notes, there are of course many languages (notably Latin and Greek) in which this grammatical illusion is not present.

conceived subject' [Anscombe 1975, p. 36]. Instead, in such acts, the subject *self-applies* the concept. Thus, 'I' is not a mysterious conception of an object, but 'signifies a form of predication' [Rödl 2007, p. 125]. The idea of a *form* of predication may be unfamiliar. That is because in analytic philosophy, we are used to thinking of predication only in terms of joining Fregean *Begriffe* to Fregean *Objekte*—predication, we are used to thinking, is just that nexus of object and concept which Frege analyzed in terms of function-application. But the current proposal, inspired by Anscombe, is precisely that our difficulties about 'I' arise because we take that Fregean nexus to be the *only* way for a subject to predicate a concept.[149] Of course, more needs to be said about what such self-predication consists in. This can be done by explaining what it is to have an unmediated agent conception of one's beliefs and actions. In §3.4.2, I elaborate on such self-predication of beliefs. In §4.3.2, I do so for the topic of self-predication of action-concepts.

So on the Anscombean view, a subject *does* predicate something of herself in first person thoughts. And in expressing such thoughts she *does* convey information about a certain object. But she predicates without first having to pick out the object in a separate act of receptive reference. In predicating something of herself in an act of self-knowledge, a subject is *thereby* put into cognitive contact with the object she predicates of. Instead of saying that 'I' is not a referring expression, it is thus more accurate to say that the contrast between reference and predication is not applicable to 'I'-thoughts. We may therefore just as well say that 'I' *does* refer to the subject, but not in the manner of receptive reference:

> One can articulate the difference of the forms of predication associated with demonstrative thought and first person thought respectively by saying that the latter does not require an act of reference, or by saying that it contains an act of reference, which therefore is no separate act. [Rödl 2007, p. 126]

In other words: if one, like Anscombe, restricts the notion of reference to receptive reference, one must say that 'I' does not refer. However, it is equally justifiable to maintain a broader notion of reference, according to which receptive reference and first-person reference are species of a common genus. Philosophers who criticize Anscombe's negative argument often fail to see that it hints at this positive conception of first-person thought. That, I think, is why they suppose that her thesis that 'I' does not refer is beyond the pale. But as I hope we can now see, denying that in first-person thought we need to cognitively reach out to an object, does not mean that such thoughts do not genuinely confer knowledge or information about the subject.

[149] For more on the idea that Fregean predication may not be the only form of predication, see Thompson [2008, pp. 13-22]. The idea that first-person thought, specifically, should not be understood as applying a concept to an object was also famously defended by Lewis [1979].

Although some philosophers may still wish to defend the idea that 'I' is a (receptively) referring expression, I think this insight may remove at least part of the motivation for doing so.

3.4.2 Reflectivism and the first person

Armed with this understanding of first person thought as self-predication, I will now argue that it follows that the reflectivist account of self-knowledge we have considered in §3.3.2 is true. For, I argue, if an unmediated conception of oneself as being *F*—where this is, e.g., believing that *p*—is self-applying the concept of being *F*, it follows that believing that *p* and knowing oneself to believe that *p* are one act. At the same time, I will explain how the reflectivist account can help us understand the nature of the kind of predication expressed in our schema '*X* self-knows being *F*'.

To start with the first lemma—that Anscombe's account of the first-person implies reflectivism about self-knowledge—let us return to the idea that 'I'-thoughts are identification-free. First-person reference (in the broader sense of 'reference' just outlined) guarantees that the one thinking the thought is the one that the thought is about. As Rödl argues, it follows that self-knowledge is 'knowledge by identity':

> . . . first person reference is a way of referring such that the object referred to is the subject referring to it, wherefore a way of knowing associated with this form of reference must fix it that an object thus known is the subject knowing it. And this it does only if knowing in this way that an object is *F* is knowing it *by being F*. Obviously, what is known in accordance with this formula includes the subject's knowledge of it. [Rödl 2007, p. 62]

Self-knowledge must be a form of knowledge such that being *F* is sufficient for knowing that one is *F*. For suppose that being *F* were not sufficient for knowledge that one is *F*. Then there would have to be some *other* relation (i.e., a relation other than identity) between the subject and her being *F* in virtue of which she comes to know that she is *F*. Her being *F* would have to influence her in some way that leads her to conclude that she is *F*. But to say that her *F*-ness would have to *affect* the (senses of the) subject is to say that the referential relation would have to be *receptive*. That is, her knowledge would have to be mediated by perception. Thus it could not be an unmediated, subjectless conception. It follows, then, that when a subject knows first-personally that she is *F*, her knowledge is identical to the thing she knows. Hence, 'what sets first person knowledge apart from sensory knowledge is that what is known first-personally *is not an independent reality of the first-person knowledge of it*' [Rödl 2007, p. 62]. For example, her believing that *p is* her self-knowing that she

believes it.[150] So the Anscombean account of first-person thought implies the falsity of the (DE) assumption: believing *p* and knowing that one believes it are *not* distinct acts of mind.[151]

As I argued in §3.3.2, it belongs to the concept of belief that a believing subject can explain *why* she believes what she does. Something is not a belief of a subject except in so far as she can show how it fits into her 'total view of reality'. So it can be no accident that in reasoning '*p*, therefore *q*', a subject comes to have knowledge of grounds (as I argued in §3.3.1). Rather, it must a be a feature of the *kind* of connection between premise and conclusion—i.e., a feature of the causality of the inference—that the one sustains the other, and that belief in the conclusion dies out when belief in the premise does so. The Anscombean account of the first person allows us to better understand the kind of causality that is constitutive of reasoning, and how it yields knowledge of grounds. To see how, let us again consider the contrast between self-knowledge and perceptual knowledge.

An act of self-knowledge differs from observational knowledge in the following way. If one comes to know that something is *F* through exercising a perceptual capacity, then the object must already be *F* before it is known to be *F*. Perceptual knowledge is a matter of registering some truth about an object which is independent of one's knowledge of it. By contrast, an act of self-knowledge, which is the same reality as what is known, does not *register* a truth about some object (that it believes that *p*, for example) but *determines*, or makes it the case, that it is true to say of the object that it believes that *p*.[152] It is, literally, the subject's determination of her views on some matter. Thus, self-predicating the belief that *p is the subject's judging that it is right to believe that p*—it is her *making up her mind*.

A subject's making up her mind is her self-predicating a thought. One way of

[150] This is why Boyle was right to say that the reflectivist account of self-knowledge is a *metaphysical* account (see §3.3.2). For according to it, self-knowledge is not just an epistemological relation—knowing that one thinks that *p* by being the subject that thinks *p*—but equally a manner of *being*, such that being *F* is knowing oneself to be *F*.

[151] Uriah Kriegel [2003] suggests that for a subject to have self-knowledge of being in some mental state is for her to be in a state that represents itself. On such a view, it seems that her self-knowledge and its object will not be two distinct acts of mind. However, as long as this idea is cashed out in terms of a propositional attitude with self-referential content, e.g., 'that *p* and I believe it', it will still not capture the insight that self-knowledge is knowledge by identity. After all, on the view Kriegel suggests, it *is* still possible to have a merely 'first-order' belief. So believing *p* will not be sufficient for knowing that one believes it. An additional step would be necessary—but as I think the argument in this chapter shows, such a step is impossible. Moreover, Kriegel's view unjustly ignores the role of self-knowledge in *constituting* belief, as I explain below.

[152] As I remarked in fn. 134, it seems that the unity formed by a thinker's thoughts must be a unity that is constituted by her consciousness of that unity. The account I develop here supports that view: for something to be a belief *is* for a thinker to have self-knowledge that she believes it. Accounts which adopt the (DE) assumption run into problems precisely because they have to appeal to some unity—the referent of 'I'—which exists prior to a subject's self-knowledge. On such a view, self-knowledge can only be accidental to the unity of the thinker.

making up one's mind is through inference. So making an inference in an act of self-predication. What Boyle called the subject's 'enduring actualization of her capacity to hold a proposition true for a reason she deems adequate' [Boyle 2011a, p. 22] *is* her act of self-predicating the concept 'believing *p*'. When someone arrives at a belief through inference, then the causality of that inference—the continuous work of the premise in sustaining the conclusion—is the subject's act of self-predicating the belief. And the other way around: self-predicating a belief that *p is* holding *p* true for reasons one deems adequate.

Now I also said that the causality of an inference must be such as to guarantee that a reasoning subject has knowledge of grounds: someone who reasons '*p*, so *q*' can *ipso facto* give a rationalization 'I believe that *q* because I believe that *p*'. We can now see how that is possible. For consider a subject who possesses knowledge of grounds that she believes *q* because she believes *p*. I argued that this explanation of her belief is a true rationalization, manifesting knowledge of grounds, only if the subject *continues* to believe that the premise supports the conclusion. Beliefs are not deposited in one's mind like a stone, and so the causal work of an inference continues for as long as the conclusion is believed. Thus our subject continues to represent the premise as a (sufficient) reason to conclude the premise, thinking '*q* is right to believe because of *p*'. But as I have just argued, the subject's thinking that *q* is right to believe in this way just is her self-predicating the concept 'believing *q*'. Therefore, her thinking that '*q* is right to believe because of *p*'—the very thought that is the causality of the inference—*is* her thinking that she believes *q* because (she believes that) *p*. So a reasoner who possesses knowledge of grounds does not perform two acts of thought: thinking first, '*q* is right to believe because . . . ', and second, 'I believe that *q* because . . . '. Rather, thinking 'I believe that *q* because I believe that *p*' *is* thinking the thought which constitutes the inference, 'it is right to believe that *q* because (it is right to believe that) *p*'.

But how is this possible? I said that someone who infers *q* from *p* thinks *both* 'it is right to believe that *q* because (it is right to believe) *p*' and *also* 'I believe that *q* because I believe that *p*'. So how can we deny that she think two *distinct* thoughts? The reason is that these phrases mutually explain each other's form. As Rödl says:

> The former phrase puts into words the form of the latter, the manner in which its elements are conjoined. It stands as '*a* is an object falling under the concept of being *F*' stands to '*Fa*'. [Rödl 2007, p. 97]

To say that someone who makes the Fregean judgment *Fa* also thinks '*a* is an object falling under the concept of being *F*' is not to say that she thinks two different thoughts. The latter phrase explicates what is implicit in the former: its form of predication. Just so, 'it is right to believe that *p*' explicates the manner of applying the

concept 'believing *p*' manifested in the unmediated first-person judgment 'I believe that *p*': namely, self-predication.

We can now understand the meaning of the slogan that self-knowledge and reasoning are one (see p. 79). It is this: reasoning *is* an unmediated, subjectless act of self-predication. And, I argue, it follows from this that an explanation of why someone believes something—of the kind that is manifested in knowledge of grounds—differs fundamentally from an explanation in terms of 'ordinary event causation'. For consider that such an 'ordinary' causal explanation of why *S* is *F* can be true independently from the subject's being able to *give* that explanation. That is, an ordinary causal explanation of why *S* is *F* can be true without *S*'s knowing that *that* is why she is *F*. By contrast, if *S* believes that *p* because she believes that *q*, where the 'because' is that of inference or reasoning, then it must be the case that *S* knows that she believes that *p* because she believes that *q*. And it is not that the truth of a belief-explanation *happens* to coincide perfectly with a subject's knowing it to be true, or that there is some further cause that ensures they coincide. Rather, a subject's reasoning '*q* because *p*' and her knowledge of grounds are not two acts of thought, but one: her self-applying the concept 'believing that *p*' *is* her judging that *p* is right to believe, *is* her concluding that *p* is true.

So the truth of a belief-explanation is *determined* by the subject's knowledge of it. There is thus no separating the causality represented by a belief-explanation from the representation of that causality.[153] Belief-explanation, then, is essentially first-personal in the following sense. In giving a (true) belief-explanation of someone else's belief, we represent that person as reasoning from premise to conclusion, saying: 'she believes that *q* because she believes that *p*'. And since her reasoning from *p* to *q* is her representing the first belief as the ground of the second, the truth of the belief-explanation is *constituted* by her thinking a first-person thought: 'I believe that *q* because I believe that *p*'.

In this way, our reflection on the nature of the link between self-knowledge and reasoning reveals that reasoning is a form of causality (and that is, a form of non-accidentality, as we have seen in §2.2.1) that cannot be reduced to ordinary event causation. It is a form of causality because an agent's reasoning explains how her *now* coming to believe that *p* depends on something else—her belief that *q*. And it is not reducible because the *way* in which the belief that *p* depends on the belief *q*—the *sense* of the 'because' in the explanation—is *through* the agent's recognition of this dependence.

[153] As Rödl [2007, p. 97] says: 'The causality of the explanation *contains* the subject's representation of this very causality' [emphasis mine]. For a more detailed exposition of the argument offered here, from which I have borrowed much, see especially his [2007, pp. 96-98].

3.5 Conclusion

It will be useful to summarize the long path we have taken. Our inquiry into self-knowledge started out by considering the phenomenon of the transparency of the question 'do I believe that *p*?' to the first-order question whether *p* (§3.1). Accounting for transparency was problematic, because, I argued, there is no valid inference from reasons or evidence in support of *p* to the conclusion that anyone at all believes it. I already suggested that it was the assumption that belief and self-belief are Distinct Existences (DE) that stands in the way of a proper account of transparent self-knowledge. In §§3.2, I reviewed Setiya's attempt to salvage transparent self-knowledge while maintaining that assumption, and clinging to an orthodox metaphysics of mind (based on propositional attitudes and event causation). I argued that his account is, despite Setiya's protestations, thoroughly deserving of the label 'reliabilism'. And as such, I argued (§3.3), it could never do justice to the phenomenon of *knowledge of grounds*, for application of Setiya's rule of transparency always has to happen behind the subject's back.

I then argued that taking the phenomenon of knowledge of grounds seriously indeed does require us to give up (DE). For the question 'Why?' that knowledge of grounds is the answer to asks for a reason that explains one's belief *presently*. So making an inference must normally put a subject in a position to know the premise for as long as she believes the conclusion. Therefore, I said, a subject must already be conscious of making an inference *in making it* (§3.3.1). Understanding how that is possible required us to reject the higher-order propositional attitude model of self-belief in favor of Boyle's reflectivism, thus rejecting (DE) (§3.3.2). As I explained, on the reflectivist account, self-knowledge is the *form* of belief: something is *not* a belief except in so far a subject represents it *as* her belief—as part of her total view of reality. Orthodox philosophy of mind obscures this by pretending that the notion of a first-order belief, or propositional attitude, is more primitive, and more lucid, than the notion of self-knowledge. To show what is wrong with this dogma of modern philosophy, we had to turn to the question of first-person reference.

As I argued, the idea that there might be belief without self-knowledge—and thus, that we can account for self-knowledge along Setiya's lines—depends on the idea that 'I' refers as a quasi-demonstrative or name. I showed how Anscombe reduces this idea to absurdity. First-person thought is immune to reference failure and misidentification, not because it is an extremely secure way of referring in the ordinary, receptive way, but because it does not refer receptively at all. The way a subject relates to herself in first-person thought is *not* through the senses, but through self-knowledge. An act of self-knowledge is thus an unmediated, subjectless act of

self-predication, which applies a concept in such a way as to *determine* that the object falls under that concept. Thus we arrived again at the reflectivist thesis: coming to believe that *p is* representing oneself as believing it.

Moreover, representing oneself as believing that *p* is judging that *p* is right to believe. A subject's coming to believe that *p* is her determining what to believe. So her representing herself as believing that *p is* her reasoning that *p* is true (for reasons *q*, *r*, . . .). Hence, I concluded, reasoning and self-knowledge are one. Reasoning is a *sui generis* form of non-accidentality, for the *causality* of reasoning—the subject's thinking that *p* is right to believe because . . . —*is* her representing herself as believing *p* on those grounds, i.e., her knowledge of grounds. There is thus no separating the causality represented by a belief-explanation, 'I believe that *p* because I believe that *q*', from the representation of that causality.

It transpires that this account of the relation between (or rather, the unity of) self-knowledge, knowledge of grounds, and inference explains what Setiya's reliabilism (or any other account that assumes (DE)) cannot explain: the truth of Evans's observation that the question 'do I believe that *p*?' is answered by attending to reasons for and against *p* itself. Reasoning about *p* leads to self-knowledge of believing *p* because reasoning to the conclusion that *p is* self-predicating the concept 'believing *p*'. Self-knowledge is thus knowledge that we do not acquire through the senses, but through reasoning. There is a traditional label for such knowledge. Knowledge that springs from the subject's thinking (as opposed to receptive knowledge, which springs from observation) is *spontaneous* knowledge. In this chapter, I have limited myself to discussion of self-knowledge of belief. But if I am right that reasoning about what is true—theoretical reasoning—signifies a form of dependence between cause and effect (premise and conclusion) fundamentally different from ordinary event causation, then it seems that this may equally apply, *mutatis mutandis*, to *practical* reasoning and its conclusion: self-knowledge of one's intentional actions. Armed with our current understanding of self-knowledge and first-person thought, that is what I will argue in the next chapter. As we will see, intentional action is another form of spontaneous knowledge—*that* is why, I will finally argue in chapter 5, it deserves the epithet 'free'.

* * *

Chapter 4

Practical knowledge

This chapter will be concerned with presenting and defending a theory of action radically different from the prevailing Causal Theory of Action (CTA). With Anscombe, I will argue that intentional action is essentially characterized by an agent's *practical knowledge* about what she is doing. In doing so, I will make heavy use of the understanding of self-knowledge and first-person thought we gained in chapter 3. Understanding practical knowledge will, at the same time, allow me to definitely refute CTA. And with the rejection of that reductionist picture of action, the way is cleared to defend, in the next chapter, the thesis that intentional action is undetermined and free action.

I will start by providing an overview of the most important tenets of Anscombean action theory (§4.1). Anscombe's account starts from the premise that agents are typically able to answer questions 'Why?' about what they are up to. Anscombe's idea is that isolating the particular sense of this question 'Why?' will be to understand intentionality: intentional actions are happenings to which a certain form of explanation is applicable. And what characterizes that question 'Why?' is precisely an agent's ability to answer it without taking recourse to inference or observation. That is, agents possess what we called *knowledge of grounds* (§3.3) of what they do intentionally. In the case of intentional action, this knowledge of grounds consists in knowledge of a certain *teleological structure*—a number of actions, or action-descriptions, being related as means to ends. This structure, Anscombe explains, is reflected in practical reasoning, which is the source of the agent's non-observational, practical knowledge of what she is doing. It would thus seem that, as I argued for theoretical reasoning, practical reasoning signifies a form of causality that is intrinsically self-conscious, and thus not reducible to ordinary event causation.

However, recent theorists of action have disputed the idea that Anscombe's insights into the nature of action and practical knowledge require us to abandon the standard causal model (§4.2). Velleman [1989] and Setiya [2007], for instance, argue that Anscombe is right that intentional action is action of which the agent has non-observational knowledge. They claim that this idea can be incorporated in variants of CTA. Thus they try to resist the idea that action-explanation is of a special, intrinsically self-conscious kind by reducing practical knowledge to a mental state that causes action in the 'ordinary', event-causal way. As I will argue, there are fundamental reasons why any view that reduces practical knowledge to event causation cannot succeed. We have already briefly encountered the phenomenon of causal deviance in §2.2.2. Here, I will show that the problem of deviance stems from the inability of CTA to properly account for the nature of an agent's representation of her own activity when acting intentionally (§4.2.3). On CTA, it is just an accident that an agent has a true representation of what she is doing that is also causally active. As we will see, this is not a bug in a particular version of CTA, but a consequence of what Lavin [2013] calls the *decompositional* nature of any causal theory, which seeks to reduce intentional action to a non-intrinsically intentional event with certain extra features (§4.3.1). Practical knowledge thus forms a fundemental and, I argue, insurmountable challenge for the very idea of a causal theory of action.

This treatment of the problem of causal deviance will show that the relation between an agent's representation of herself and the action which she thus represents cannot be one of ordinary event causation. Rather, an agent's practical representation of her φ'ing must settle it that she *is* φ'ing. Practical knowledge, I argue (§4.3.2), is thus a species of unmediated self-predication, which we have already encountered in §4.3.3. On the Anscombean view, practical knowledge is a species of self-knowledge—knowledge that is 'the cause of what it understands' [Anscombe 1963, p. 87]. Such knowledge *is* the conclusion of practical reasoning, i.e., an agent's answer to the question what to do, and is therefore *identical to the agent's action*. Anscombe's theory thus provides a radical alternative to the decompositional orthodoxy. In §§4.3.3-4.3.4, I further explain the relation between 'mere' events and intentional actions, and how the Anscombean theory can accommodate both the physical and the rational aspects of action.

4.1 Intention and the question 'Why?'

Anscombe's project in *Intention* is to elucidate the concept of 'intention' under three headings: expression of intention, intention for the future, and intentional action [Anscombe 1963, p. 1]. She attempts to account for the unity of these three topics,

showing how a single concept of intention is at work in all of them. In this chapter, I will mainly focus on her account of intentional action, only briefly touching the topics of expression of intention and intention for the future. This is in line with Anscombe's own methodology: she insists that understanding intentional action is fundamental for understanding the other two.[154]

What, then, is the difference between an intentional action and a mere happening or bodily movement? As we have seen (§1.2), CTA attempts to answer this question by taking seriously Wittgenstein's question: 'what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?' [Wittgenstein 2001 [1953], §621] Anscombe's approach is different, rejecting that question (as her teacher would surely have approved). She rather tries to isolate a special form of explanation that distinguishes actions that are intentional from those that are not. Intentional actions are 'actions to which a certain sense of the question 'Why?' is given application' [Anscombe 1963, p. 9]. And it is easy to give a name to that sense of 'Why', but not so easy to give a philosophical account of it:

> . . . the sense is of course that in which the answer, if positive, gives a reason for acting. But this is not a sufficient statement, because the question 'Why?' and 'What is meant by reason for acting?' are one and the same. [Anscombe 1963, p. 9]

An account of the question 'Why?' *is* an account of reasons for action. And we cannot elucidate the relevant notion of a reason by insisting that it is a reason *for action*:

> 'Giving a sudden start', some might say, 'is not *acting* in the sense suggested by the expression 'reason for acting' [. . .]' Why is giving a start or a gasp not an 'action', while sending for a taxi, or crossing the road, is one? The answer cannot be 'Because the answer to the question 'why?' may give a *reason* in the latter cases', for the answer may 'give a reason' in the former cases too; and we cannot say 'Ah, but not a reason for *acting*'; we should be going round in circles. [Anscombe 1963, p. 10]

Answering the question 'why did she give a sudden start?' by saying, e.g., 'because she is easily frightened' is giving a *kind* of reason, but not the kind applicable to intentional action. Understanding *that* sense of the question 'Why?' requires a non-circular definition:

> We need to find the difference between the two kinds of 'reason' without talking about 'acting'; and if we do, perhaps we shall discover what is meant by 'acting' when it is said with this special emphasis. [Anscombe 1963, p. 10]

There are, thus, multiple senses of the question 'Why?', or as we have called them, multiple forms of explanation. Anscombe's challenge is to account for the sense of

[154] In fact, after the long stretch of the book that is involved only with intentional action, Anscombe [1963, p. 90] returns to 'expression of intention for the future', remarking only that 'What I have said about intention in acting applies also to intention in a proposed action'.

'Why?' (or form of explanation) associated with intentional action, without taking recourse to notions that already take some account of it for granted. How could we do this? Anscombe's approach is to first delineate the class of actions to which the question 'Why?' (which we shall henceforth use to refer to the question pertaining to intentional action) negatively, showing when the question is *not* applicable. In a further step, as we will see, she tries to give a positive account of the question's sense.

The question 'Why are you doing φ?' is most clearly refused application when an agent does not know that she is φ'ing. This seems natural: imagine someone who is asked 'Why are you ringing that bell?' and replies 'Good heavens! I didn't know *I* was ringing it!' [Anscombe 1963, §28] (the agent was, e.g., unwittingly leaning against the doorbell). In such a case, the subject refuses the demand for reasons for the (putative) action. And if she is sincere, i.e., she really did not know that she was ringing the bell, then the question really was misplaced, precisely because the agent does not know how to answer it.[155] Interestingly, the question is sometimes also refused application when the agent *does* know that she is performing it, but where her knowledge is of the wrong kind:

> Say I go over to the window and open it. Someone who hears me moving calls out: What are you doing to make that noise? I reply 'Opening the window'... But I don't say the words like this: 'Let me see, what is this body bringing about? Ah yes! the opening of the window'. [Anscombe 1963, p. 51]

That is, the question 'Why?' is equally refused application when the subject has to look and see what she is doing—i.e., when her knowledge of what she is doing is merely *observational* [Anscombe 1963, p. 14].[156] Now it is important to note that, for Anscombe as for Davidson, actions are only intentional under a description [Anscombe 1963, p. 12].[157] For example, if an agent is sawing a plank, thereby creating some sawdust, it may be that she is intentionally sawing, but not intentionally creating sawdust. Yet both descriptions, 'sawing a plank' and 'creating sawdust', are descriptions of the same action, in the sense that she does not have to do anything *other* than saw a plank for it to be true that she is creating sawdust. Hence, 'the statement that a man knows he is doing X does not imply the statement that, concerning anything which is also his doing X, he knows that he is doing that thing' [Anscombe 1963, p. 12]. And if someone knows without observation that she is sawing a plank, but knows only by observation that she is also creating sawdust, the question 'Why?'

[155] As Wittgenstein says, intentional action is 'marked by the absence of surprise' [Wittgenstein 2001 [1953], §628].

[156] Anscombe further argues that the question 'Why?' is also refused application when the answer is a 'mental cause' (e.g., 'I jumped because I saw a scary face in the window') or a 'motive' [Anscombe 1963, §§10-13]. The latter category will not concern us here, but see §4.2 for discussion about the (potential) role of mental causes in answering the question 'Why?'.

[157] See also [Anscombe 1979].

is applicable only to the former, and not to the latter, description of what she is doing.

The class of intentional actions is thus a subclass of the class of things known *without* observation. As a paradigmatic example of other members of that class, Anscombe cites the ability to know the position of one's limbs. Some interpreters have taken this to suggest that the knowledge an agent has of her intentional actions is the same kind of knowledge she has of the position of her limbs. And some [Pickard 2004, e.g.] have disputed the whole idea that agents know what they are doing without observation because they argue, further, that knowledge of the position of one's limbs *is*, after all, observational—*proprioception* being a form of perception. Even if it is true that knowledge of the position of one's limbs does not fall in the class of things known without observation, however, it would not follow that knowledge of one's intentional actions is similarly observational: non-observational knowledge, according to Anscombe, is a class of which knowledge of one's intentional actions is a subclass. So whether Anscombe is right about other putative subclasses is besides the point when it comes to her account of action. For the kind of knowledge one has of one's intentional actions is, as we will now see, *not* the same as knowledge of the position of one's limbs.

That this is so is apparent from the fact that knowledge of one's intentional action is not exhausted by knowledge *that* one is doing something, but also includes one's reason for doing it.[158] That is, when the question 'Why are you φ'ing?' is not refused application, then an agent will also have knowledge without observation of the *answer* to that question. For an answer to the question 'Why?', Anscombe observes, is typically something the agent intends to do: one might, for example, answer the question 'Why are you sawing a plank?' by saying 'because I'm building a house'. If one's building a house is something one intends to do, of which one's sawing a plank is the means, then it follows that building a house must be a description of one's action that is also known without observation. Moreover, it seems that one's non-observational knowledge is not limited to the two description's of one's action—'sawing a plank', 'building a house'—but also includes the *explanatory connection* between the two. One knows, without having to look and see, that one is sawing a plank *because* one is building a house. It seems impossible to know *that* through a kind of proprioceptive experience.

The specific kind of knowledge one has of what one is doing intentionally—what Anscombe calls 'knowledge in intention' or *practical* knowledge—is not knowledge of just one action, but of an entire chain of actions, connected as means to ends. Consider Anscombe's [1963, pp. 37-41] famous example of a gardener pumping

[158] This can be put by saying that knowledge of the position of one's limbs is an example of what Anscombe, in her essay on the first person, called an unmediated *patient* conception, while knowledge of what one is doing intentionally is, as I will argue (§4.3.2), an unmediated *agent* conception.

water:

- 'Why are you moving your arms?' (*A*) – 'I'm pumping water' (*B*)
- 'Why are you pumping water?' (*B*) – 'I'm replenishing the water supply' (*C*)
- 'Why are you replenishing the water supply?' (*C*) – 'I'm poisoning the inhabitants' (*D*)

Each of *A*-*D* are descriptions under which the agent's action is intentional: she knows without observation that she is doing *A*, *B*, *C*, and *D*. Now an action is intentional *in virtue* of the fact that the question 'Why?' applies to it, and the question 'Why?' applies if and only if one has non-observational knowledge of the answer to it. So each description, up to the last one, is a description under which the gardener's action is intentional *in virtue* of the fact that she knows that she is doing it *because* she is doing the next thing in the chain. For example, doing *A* is an intentional description in virtue of the fact that the gardener knows: 'I'm doing *A* because I am doing *B*'. But her doing *B* is only a fitting answer to the question 'Why are you doing *A*?' insofar as it is intentional. And it is intentional because the gardener knows: 'I'm doing *B* because I am doing *C*'. So in the end, the agent's non-observational knowledge of what she is doing includes the whole chain, or the whole *A*-*D* order, as Anscombe calls it: it is knowledge that she is doing *A* because she is doing *B*, because she is doing *C*, because she is doing *D*. So any of the higher-level descriptions will do equally well as an answer to the question 'Why are you doing *A*': it is as true to say that the gardener is doing *A* because she is doing *D*, as it is to say that she is doing *A* because she is doing *B* or *C* [Anscombe 1963, p. 46].[159]

The mark of the kind of non-observational knowledge of one's intentional actions, as opposed to the broader class of things known without observation, is thus that it is knowledge of how different descriptions of what one is doing intentionally relate as means to ends. In the gardener example, we are dealing with four temporally coinciding descriptions: supposing the poison is already in the water, the gardener is already doing all he has to do in order to execute his final intention. But this is an artifact of the example. Suppose that an agent is chopping some onions, because she is making a sauce, because she is making dinner. Of course, making a sauce requires more than chopping onions, so the agent will have more work to do after the last onion is chopped. Nevertheless, we find the same kind of means-ends structure as in the gardening example: the agent is (*A*) moving her hands, because she is (*B*) chopping some onions, because she is (*C*) making a sauce, because she is (*D*) making

[159] What about the final link in the chain, *D*? In virtue of what is that a description of the agent's action under which it is intentional? As Anscombe argues, all means-ends chains end in description of a particular kind—what she calls a desirability characterization. I explain this in §4.1.1 below.

dinner. Chopping onions is a *phase* or a *part* of the larger action of making dinner, just as pumping is a (temporally completely overlapping) part of replenishing the water supply. This kind of structure, Anscombe argues, is characteristic of all intentional action:[160] it is 'an order which is there whenever actions are done with intentions' [Anscombe 1963, p. 80].

So far, I have described the sense of the question 'Why are you φ'ing?' negatively, by pointing out that the question is *not* applicable when the agent lacks non-observational knowledge of her φ'ing, or of whatever she is doing φ for. Positively, I have described the means-ends structure between different intentional descriptions of an action that is revealed by asking a series of 'Why?' questions. But that is not yet to give a philosophical account of the question's sense. Since the question 'Why?' is applicable if and only if the agent has a certain kind of non-observational knowledge, a positive philosophical account will have to explain the nature of this *practical* knowledge. How can an agent come to know what she is doing, and why, if not through observation?

As Michael Thompson points out in his discussion of the nature of practical knowledge, what makes it such knowledge so difficult to understand is that its object (an action) is equally something that we can observe:

> Anscombe insists everywhere that an agent's intentional action is just another sort of process in the world, something perceptible and watchable by others. Indeed her first attempt to isolate the category of intentional action, in a section that might easily be overlooked, depends on this. She says that if I ask you to enter a room with some people in it and to report what's up with them, the process-descriptions you return with will mostly be descriptions of intentional actions.[161] [Thompson 2011, p. 201]

What someone is doing intentionally, and why, *is* often a perfectly perceptible fact—although of course, there may be cases in which it's a little more difficult to tell what someone is up to, and we may have to resort to actually asking questions 'Why?'. And this makes for the following difficulty:

[160] The idea that this means-ends structure is, for Anscombe, what characterizes all intentional action, as well as what unifies the three subjects of intentional action, intention for the future, and expression of intention, is defended in [Moran and Stone 2011a]. Note that Anscombe does not deny that one may also intentionally perform actions for no *further* aim—for instance, one may, e.g., take a stroll or whistle a tune 'for no particular reason' [Anscombe 1963, p. 25]. The thought is that these are limiting cases of actions with means-ends structure: they are cases of doing something for its own sake.

[161] Thompson is referring to the following passage:

> . . . in a very large number of cases, your selection from the immense variety of true statements [. . .] which you might make would coincide with what [the agent] could say he was doing [. . .] without adverting to observation. I am sitting in a chair writing, and anyone grown to the age of reason in the same world would know this as soon as he saw me. . . [Anscombe 1963, p. 8]

> Now if there are *two* ways of knowing here, one of which I call knowledge of one's intentional action and the other of which I call knowledge by observation of what takes place, then must there not be two *objects* of knowledge? How can one speak of two different knowledges of *exactly* the same thing? [Anscombe 1963, p. 51]

If what is known practically is something that it is impossible for the agent to know merely by observation, must it not be something entirely interior, e.g., a mental act of *willing*? True to her Wittgensteinian heritage, Anscombe dismisses any such story of purely inner acts as 'a mad account' [Anscombe 1963, p. 52]. We cannot separate the intentional *doing* of an act, which on this picture is private and known practically, from what *happens*, which would be public and known by observation. Instead, Anscombe formulates the following dictum about what is known practically:

> I *do* what *happens*. That is to say, when the description of what happens is the very thing which I should say I was doing, then there is no distinction between my doing and the thing's happening. [Anscombe 1963, pp. 52-53]

So the reason why practical knowledge is 'without observation' is not that what is known practically is, in principle, hidden from sight, as a purely mental, dualist act of will might be. But how can there be non-observational knowledge of what is an actual, real-world happening? Many commentators have declared that such knowledge would be impossible. They insist that we may, of course, know without observation what we intend, but that it would be a mere 'licensed wishful thinking' [Grice 1971, pp. 263-279] to suppose that what we intend to do corresponds to what is actually happening. After all, we frequently fail to execute our intentions. We will return to these cases of failed action in §4.3.4. For now, let us set such skepticism aside, and focus on how Anscombe tries to account for practical knowledge of what is happening. And as Anscombe says: 'The notion of 'practical knowledge' can only be understood if we first understand 'practical reasoning' ' [Anscombe 1963, §33].

4.1.1 Practical reasoning

Anscombe's account of practical reasoning is heavily based on Aristotle's. The best way to understand it, and to see in which way it differs from other accounts of practical reasoning we have already seen, will thus be to take as our example a practical syllogism inspired by Aristotle:

> Dry food is wholesome for humans.
> This cracker is a piece of dry food.
> ∴ So I'll have the cracker.

This syllogism has some features that sharply distinguish it from the Davidsonian conception of practical reasoning we have seen in §1.2, according to which an agent

weighs the strength of various pro-attitudes against each other. To start with, the first 'major' or 'universal' premise contains a 'desirability characterization' [Anscombe 1963, §36,71-72], 'wholesome', and a kind-term, 'human'. Anscombe suggests in *Intention* that other kind-terms ('farmer', 'philosopher') could in principle be substituted.[162] But should we not, then, add a further premise, 'I am human', or 'I'm a philosopher'? It seems not. This is because 'I am human' does not seem to be analogous to the belief that 'this cracker is a piece of dry food': it is not one of the *circumstances* one finds oneself in, to be taken account of in deciding how to gather dry food. For without the recognition that the major premises are relevant to (such as) *oneself*, no reasoning about what to do would get off the ground in the first place. Indeed, in some examples, Aristotle (and, following him, Anscombe) omit the kind-term in the major premise—e.g., 'Anything sweet is pleasant' [Anscombe 1963, p. 64]. The reference to the kind-term is taken to be implicit.

Secondly, Anscombe insists the conclusion should be thought of as an action, for which the form of words 'So I'll...' only stands proxy on paper. In fact, we must take seriously the thought that the above syllogism is, in a very important sense, *not a practical syllogism at all*: it is an 'idle practical syllogism which is just a classroom example' [Anscombe 1963, p. 60]. By contrast, in the 'practical syllogism proper', 'the conclusion is an action whose point is shewn by the premises, which are now, so to speak, on active service' [Anscombe 1963, p. 60]. Anscombe thus insists that there is a fundamental difference between reasoning theoretically to the truth of a conclusion, and practical reasoning, leading to an action.

The idea that there is such a specifically practical form of reasoning, concluding in action and differing in kind from theoretical reasoning, is obviously disputed. For now, notice a third surprising feature of Anscombe's account, as compared to e.g. Davidson's rendering of a practical argument: the absence of the agent's desires. Where for Davidson these play an essential role in both the premises and the conclusion (which might be explicated as, e.g., 'φ'ing is what I desire most'), Anscombe's syllogism does not mention the fact that a particular thing is *wanted* at all. Nor is this an artifact of our particular example. It is simply *not* the case, Anscombe insists, that the premises *could* include desire ('dry food is what I desire', etc.):

> . . . the role of 'wanting' in the practical syllogism is quite different from that of a premise. It is that whatever is described in the proposition that is the starting-point of the argument must be wanted in order for the reasoning to lead to any action. [. . .] 'Dry food . . . suits anyone etc., so I'll have some of this' is a piece of practical reasoning

[162] Anscombe later [1969] retracts this suggestion, arguing, in effect, that all (human) practical reasoning departs from a major premise running '*X* is good for humans'. This question need not concern us here.

> which will go on only in someone who wants to eat suitable food. [Anscombe 1963, p. 66]

Anscombe likens the role of wanting in practical reasoning to that of belief in the theoretical domain. When one reasons theoretically to the truth of q from a premise p, the fact that one *believes* that p is, of course, not part of the *content* of the premise.[163] But such an inference will not occur in one who does not believe that p. So, in the case of action, a practical syllogism like the one above will only go on, or be pursued up to the conclusion, by someone who wants wholesome dry food. We might thus say that desire is the driving force behind practical reasoning, as belief is the driving force behind theoretical reasoning.

This point requires some more detailed explanation. Is it not true that one can do something intentionally because one desires something? For instance, someone might go to the supermarket because he has a particularly strong craving for something sweet—chocolate, say. Anscombe admits that, of course, this is possible, and in such cases desire really is part of the major premise. But we are then talking about a quite specific notion of 'desire', namely, desire in the sense of the anticipation of the pleasure of some prospective action. This notion of desire or wanting need not be the one that underlies action: when one acts intentionally, one *wants* to achieve something, but that doesn't mean that the action is aimed at pleasure. Indeed, that something is or would be pleasurable is just one possible *content* of a major premise: 'Eating something sweet is pleasurable; chocolate is sweet; so let me eat some chocolate; they have chocolate in the supermarket—so let me go there'. But, Anscombe insists, it would be foolish to insist that *all* practical inference proceeds from major premises of the form 'doing X is pleasurable', as e.g. the classical utilitarians did: 'They were saying that something which they thought of as like a particular tickle or itch was quite obviously the point of doing anything whatever' [Anscombe 1963, p. 77]. On the face of it, there are quite many things that might be the point of an action, and thus figure in the major premise of a practical inference: e.g., health, safety, and perhaps even friendship and justice.

Thus, if desire plays a role in practical reasoning generally, it cannot be the sense of desire associated with pleasure.[164] Anscombe explains elsewhere [1974] that we *can* think of the major premise in a practical inference as transmitting the desirability of one action (eating something sweet) to the next (eating some chocolate), all the way to a conclusion (going to the supermarket) that one can perform without needing to

[163] In the previous chapter, we have argued that one's self-knowledge of the premise is rather the *form* of the premise.

[164] It will rather be desire in the broader sense of a faculty of appetite, or of willing or striving to get something. Compare Thompson [2008, pp. 103-104].

deliberate any further—'until straightaway he acts', as Aristotle says.[165] But we can see that what is transmitted from each step is obviously not the pleasure of eating something sweet, or anticipation of the delicious taste of the chocolate. It is not 'wanting' in *that* sense that drives practical reasoning. Rather, the kind of wanting at issue is, as I said, analogous to belief in theoretical reasoning. In a theoretical inference, what is transmitted from premise to conclusion is that the latter is *to be believed* in light of the first. In a practical inference, what is transmitted from one step to the next is that the one is *to be done* in light of the other. To say that a premise or conclusion of practical reasoning is something one *wants* or *desires* is thus not to say that it is something one feels a thirst or craving to do, but that it is something one judges as *good to do*:

> The conceptual connexion between 'wanting' [. . .] and 'good' can be compared to the conceptual connexion between 'judgement' [i.e., belief] and 'truth'. Truth is the object of judgement, and good the object of wanting. . . [Anscombe 1963, p. 76]

Anscombe is careful not to give too substantive, or ethical, a content to the notion of 'good' employed here: all she means is that in a practical inference, 'the premises shew what good, what use, the action is' [Anscombe 1974, p. 114]. Wanting something is judging that it is good to do in the sense of contributing to some goal.[166] The kind of 'wanting' that drives practical reasoning is thus closely connected to the question 'Why?': what someone *wants*, in this sense, is something he is *after*.[167] Therefore, Anscombe argues, an account of practical reasoning as reasoning *about* desires, 'is as incorrect as it would be to represent theoretical reasoning in terms of belief' [Anscombe 1974, p. 138]. She grants that there may be true entailments of the form 'if X believes that p and that $p \to q$, then he believes that q'.[168] However,

> We would never think that the validity of 'p, if p then q, therefore q' was to be expounded as the entailment of 'X believes q' by 'X believes that p and that if p then q'. It is, we feel, the other way around. [Anscombe 1974, p. 138]

Similarly, we should not analyze practical reasoning in terms of propositions about someone's desires, that jointly entail that she performs, or will perform, some action. The premises of a practical syllogism do not logically entail or necessitate the occurrence of the concluding action. Rather, when there *is* an intentional action, there will be some practical syllogism that shows what the point of it is in the agent's eyes.

[165] See *On the motion of animals*, 701a10-15.

[166] This is not to say that the notion of goodness relevant to practical reasoning is irrelevant to ethics. As Anscombe [1969] argues, the fact that practical reasoning is aimed at the good means that it presupposes the idea of a form of life (or a lifeform), which sets a normative standards for the individuals who fall under it. Compare my discussion of lifeforms in §5.3.1.

[167] Anscombe [1963, p. 68] remarks: 'The primitive sign of wanting is *trying to get*'.

[168] These entailments hold, on her view, just because we would not count someone who does not accept such an entailment as really believing that p [Anscombe 1974, p. 130].

And of course, such practical reasoning does not have to occur explicitly or 'before the mind' [Anscombe 1963, p. 79]:

> Generally speaking, it would be very rare for a person to go through all the steps of a piece of practical reasoning as set out in conformity with Aristotle's models, saying e.g. 'I am human', and 'Lying on a bed is a good way of resting'. [...] The interest of the account is that it describes an order that is there whenever actions are done with intentions; the same order as I arrived at in discussing what 'the intentional action' was, when the man was pumping water. [Anscombe 1963, pp. 79-80]

Thus, the order of practical reasoning *is* the *A*-*D* order, or the means-ends structure, that is revealed by the question 'Why?'. Indeed, we can see that each step up the chain of answers to the 'Why?'-question corresponds to a step down in the ladder of practical reasoning: doing *B* (pumping water) is the answer the question *why* one is doing *A* (moving one's arm), and doing *A* (moving one's arms) is the answer to the question of practical reasoning, namely, of *how* to do *B* (pump water). Thus, an action-explanation 'I am doing *A* because I am doing *B*' represents *the same dependence between A and B* as an agent's reasoning practically 'I'm doing *B*, so let me do *A*'—looked at from a different perspective. In the next section (§4.1.2), we will see how this identity can help us understand practical knowledge, as Anscombe promised.

4.1.2 Practical reasoning and practical knowledge

In order to isolate the particular sense of the question 'Why?' that applies to intentional action, I said it would be necessary to give an account of practical knowledge, which allows an agent to answer that question. The difficulty we encountered in accounting for that knowledge was that it is supposed to be non-observational knowledge of what is *happening*—i.e., of a public, observable process. How can understanding practical reasoning help us to understand how *that* is possible, as Anscombe promised?

In the paragraph in which she herself returns to the topic of practical knowledge, Anscombe gives the following analogy:

> Imagine someone directing a project, like the erection of a building which he cannot see and not get reports on, purely by giving orders. His imagination (evidently a superhuman one) takes the place of the perception that would ordinarily be employed by the director of such a project. He is not like a man merely considering speculatively how a thing might be done; such a man can leave many points unsettled, but this man must settle everything in *a* right order. *His* knowledge of what is done is practical knowledge. [Anscombe 1963, p. 82]

The director is in a position to say what is happening (that a building of such-and-

such specifications is getting erected) because he has reasoned practically about how the project should proceed. His planning of the building steps allows him to do without the perception of the project that someone who is *not* the director of the project, but a mere onlooker, would require to know what is going on—we might imagine the spectator exclaiming 'ah, so it's going to be one of these modern office buildings!', when he sees the builders putting the glass facade into place, as the director had planned from the beginning. And indeed, it intuitively seems that being reduced to spectatorship—literally, having to look to see what's going on—is the very opposite of exercising agency: 'the essence of passivity with respect to an event is witnessing it' [1989, p. xiii]. So it seems that Anscombe is saying that the link between practical reasoning and practical knowledge is that the latter is 'maker's knowledge' [Velleman 2006b, p. 262]. When we act intentionally, we are in the same position with respect to what we are doing as the director is with respect to the building.

Many contemporary philosophers of action—even those working in the tradition of CTA—treat the idea that practical knowledge is maker's knowledge, and that this distinguishes agency from spectatorship, as Anscombe's most important insight.[169] For example, here is Velleman:

> The designer of something is the one whose conception of the thing determines how it is, rather than *vice versa*, and determines this not by chance but a mechanism reliable enough to justify his confidence in that conception as an accurate representation of the thing. To be the designer of something is just to be the one whose conception of it has epistemic authority by virtue of being its cause rather than its concomitant or effect. [Velleman 2005, p. 227]

Velleman says that practical knowledge has epistemic authority by being the cause, rather than the effect, of the happening it is about. Indeed, something like this is also Anscombe's view:

> . . . the account given by Aquinas of the nature of practical knowledge holds: Practical knowledge is 'the cause of what it understands', unlike 'speculative' knowledge, which 'is derived from the objects known'. [Anscombe 1963, p. 87]

Practical reasoning and practical knowledge thus have a distinctive relation to their object, the agent's intentional action. Theoretical or speculative knowledge, e.g., knowledge acquired through perception, is a representation caused by what it represents—the object known causes the representation through affecting the subject's senses. By contrast, practical knowledge—which is the conclusion of the agent's practical reasoning, '. . . so I'll φ'—is a representation that is *productive* of what it rep-

[169] The idea that in reasoning practically, we design our own actions has thus found its way into sophisticated modern versions of CTA. I will examine these theories, and argue that they cannot do justice to Anscombe's insight, in §4.2.

resents.[170]

Still, one might wonder, how does the idea that practical knowledge is the cause of what it understands help us with the epistemological problem? We wanted to explain how practical knowledge could be non-observational knowledge of what is happening, not just of the agent's willing or intending to do something. And can we not fail to do something that we, through practical reasoning, have decided we ought to do? If so, how can we be confident and justified enough to say that we *are* φ'ing, just on the basis of practical reasoning? According to Velleman, Anscombe is not worried about this:

> Anscombe was also, as I interpret her, a reliabilist about knowledge—in particular, about what is 'known by being the content of intention'. She thought that a reliable connection in general between what's intended and what's done is sufficient to confer the status of knowledge on a particular intention, provided that the connection holds up in the particular case. [Velleman 2005, p. 228]

As long as an agent's practical representation and his action are linked together by 'a mechanism reliable enough to justify his confidence' [Velleman 2005, p. 227], the agent's representation of what he is doing may still count as knowledge, 'provided that the connection holds up in the particular case'. Velleman denounces what he says is Anscombe's reliabilism about knowledge, and expresses his hope 'that knowledge without observation can meet the justificatory standards of internalist epistemology' [Velleman 2005, p. 229]—although he adds that he thinks the dispute is not especially significant.

However, the idea that practical reasoning is a mechanism that reliably produces the action represented in its conclusion is unlikely as an interpretation of Anscombe.[171] For if that was how Anscombe viewed the link between practical reasoning and practical knowledge, it would become quite unclear how that would help us to understand the special sense of the question 'Why?'. Just saying that there is a reliable connection between an agent's drawing a conclusion in practical reasoning and her actually performing the action she thereby derived does nothing to illuminate the sense of the 'therefore' that connects the steps in the agent's reasoning. And without an explanation of this 'therefore', we remain in the dark about its converse: the 'because' that answers the question 'Why?'. Of course, Velleman, as a supporter of CTA, supposes that the sense of the 'because' of action-explanation is simply that

[170] If one wishes to reserve the term 'representation' for speculative knowledge, then one might want to adopt Anscombe's translation of Aquinas, and say that practical knowledge is productive of what it 'understands'.

[171] It is unfortunate that Anscombe's essay on the first person (see §3.4) is so often treated as irrelevant to her account in *Intention*. If practical knowledge were recognized as a species of the unmediated agent conceptions which we have seen in §3.4.1, it would be immediately obvious that ascribing any form of reliabilism to her is absurd.

of ordinary causal explanation. But Anscombe, as we have seen, thinks we cannot take the sense of the question 'Why?' for granted like that. So her claim that practical knowledge is 'the cause of what it understands' must mean something different than that practical reasoning is a mechanism that reliably produces intentional actions.

Then in what way *does* practical reasoning lead to the acquisition of practical knowledge, on the Anscombean view? The answer, I suggest, lies in understanding that practical knowledge is a species of *self*-knowledge. In fact, the strong parallel between self-knowledge of belief (of which we treated in the previous chapter) and practical knowledge should already be emerging. For practical knowledge shares a number of important characteristics with its theoretical sibling. First of all, it is essentially *first personal* knowledge: an agent will only be able to answer the question 'Why?' if she thinks, first personally, '*I* am doing φ'. Second, as in the theoretical case, practical knowledge is not limited to knowledge *what*, but also includes knowledge of grounds: the agent knows both *that* and *why* she is doing φ. Third, we have seen that this knowledge of grounds (in the theoretical case) is knowledge of the reasons for which one *presently* believes that p: it is knowledge that 'I believe that p because I believe that q'. Similarly, an intentionally acting agent knows 'I am doing φ because I am doing ψ'—where the latter will remain true throughout the time she is φ'ing.[172] And finally, practical knowledge seems to be *transparent* in a way analogous to self-knowledge of belief: one comes to know that one is doing φ by reasoning practically about what to do, just as one comes to self-know that p by reasoning theoretically about what to believe.

Now it may seem that there is an important disanalogy between theoretical and practical self-knowledge. In the previous chapter, I argued against the popular idea that one's belief that p and the knowledge that one has this belief are distinct existences—the (DE) assumption. But in the practical case, the objection goes, such a rejection of (DE) would be impossible. For again, what is known practically is something that happens in the external world, an event, say a sinking of the Bismarck or a replenishing of the water supply. It is definitely not a mental state, and therefore cannot be the same reality as the knowledge of it.

However, I will argue in the remainder of this chapter that the (DE) assumption is as false in the case of practical knowledge as it is in the theoretical domain. As we will see, that is why Anscombe suggests that we must reject a purely 'contemplative' [Anscombe 1963, p. 57] conception of knowledge. Thus the relation between practical knowledge and its object—an agent's intentional action—is analogous to what I argued is the relation between (theoretical) self-knowledge and belief. Believing

[172] Except in rare cases in which the agent is first doing φ because she is doing ψ, then drops the intention to do ψ, but keeps φ'ing for some different reason.

that p, I argued, *is* self-predicating the concept of believing that p. Similarly, acting intentionally is self-predicating an action-concept, 'doing φ'. Acting intentionally, then, *is* self-knowing that one is doing φ.

I will argue for this view by considering theories that attempt to account for practical self-knowledge while upholding the practical analogue of the (DE) assumption. These theories argue that an intention is a belief-state (i.e., as CTA conceives of beliefs). I will argue that these attempts to account for practical knowledge on reductionist lines fail. Furthermore, I argue, this shows that CTA is a fundamentally flawed theory of action. I will then develop the idea that practical knowledge should be thought of as self-predication of an action concept in more detail. In §4.3.4, we will see how this allows Anscombe to deal with the epistemological problem formed by the possibility of failed execution of intention, without espousing reliabilism.

4.2 Intention as belief

In this section, I will scrutinize the idea that practical knowledge can be accounted for as a belief-like state of the agent, whose contents represent her as doing φ (or as *going* to do φ). The notion of belief at play here is the notion of belief as a mental state (or mental event), as it figures in broadly Davidsonian theories. The theories I will examine are thus attempts to accommodate Anscombe's insight that there is a strong link between intentional action and an agent's knowledge thereof within the framework of CTA. Although the idea that intention is a belief or a belief-like mental state is fairly widespread,[173] I will take as paradigmatic the most well known, and arguably most well-developed intention-as-belief accounts: those of Velleman (§4.2.1) and Setiya (§4.2.2).

4.2.1 Velleman: the desire for self-understanding

Velleman's account of the relation between action and self-knowledge departs from the idea that knowledge about what one is doing is a necessary condition of acting intentionally:

> You are walking up Fifth Avenue. All of a sudden you realize that you don't know what you're doing. You can see that you're walking up Fifth Avenue, of course: the surroundings are quite familiar. But the reason why you're walking up Fifth Avenue escapes you, and so you still don't know what you're doing. Are you walking home from work? Trying to catch a downtown bus? Just taking a stroll? You stop to think. [Velleman 1989, p. 15]

[173] See, for example, Harman [1997].

Without knowledge of what we're doing in walking up Fifth Avenue, our action comes to a halt. And this knowledge cannot be observational: perception does not reveal under which description of one's action it is intentional. So, Velleman rightly supposes, it is not merely an interesting but superficial fact about intentional action that it is often accompanied by knowledge of what we're doing. The representation of one's action plays a *productive* role in bringing that action about. Acting intentionally *is* acting in such a way as to make one's representation of what one is doing true, culminating in self-knowledge. Velleman defines intentions as 'self-fulfilling expectations that are motivated by a desire for their fulfillment and that represent themselves as such' [Velleman 1989, p. 109]. That is, an agent starts with certain desires, e.g., to φ, and on that basis forms a belief *that he will φ on that very basis*. Or more explicitly, the agent thinks:

> "Because I have these motives for getting myself to do this (and I know it), I'm hereby reinforcing those predispositions to the point where I'll do it next." [Velleman 1989, p. 88]

The idea is thus that an intention is, literally, a self-fulfilling prophecy. But how does the belief 'reinforce' the preexisting desires in such a way as to justify the agent's expectation that he will actually perform the action? Is belief not, as Humeans would have us think, impotent to influence our conduct? Velleman argues that an intention, as he conceives of it, *is* able to guide one's conduct because we all have a standing desire to know what we're doing. This desire for self-knowledge combines with the belief (intention) that one will φ, causing one's action of φ'ing, and thus rendering the belief true. Velleman explains that this 'general tendency' to know oneself is thus an 'intellectual passion' [Velleman 2005, p. 229], a motive in the service of reason. It is only because of this general tendency to understand our conduct—what we're doing and why—that intentional action is possible at all. According to Velleman, therefore, 'self-knowledge is the constitutive aim of action' [Velleman 2000, p. 26].

It is important to note that Velleman's accounts for the interaction between intentions and the desire to know what we're doing is strictly causal (i.e., ordinary event causation) terms. Velleman claims that this is a virtue of his account over Anscombe's, who leaves us in the dark about how practical reasoning leads to action *and* practical knowledge:

> I attempt to explain how a mental state that causes its object can still qualify as knowledge in a reasonably familiar sense, whereas Anscombe's account of an agent's self-knowledge leaves it looking not just causally perverse but epistemically mysterious. [Velleman 1989, p. 103]

As I will argue later, Anscombe's account is not 'causally perverse', nor does she lack an account of the epistemic credentials of practical knowledge. But how epistemically

sound is Velleman's own theory? As he himself puts the worry:

> For how could you ever have grounds for expecting an action whose very occurrence would remain unlikely until you expected it? You would never have justification for forming this expectation unless you'd already formed it, and so you could never justifiedly form it, in the first place. [Velleman 1989, p. 56]

The solution, for Velleman, is that *forming* the expectation (i.e., the intention) is epistemically permissible because an agent knows that she has a tendency to do what she expects [Velleman 1989, pp. 62-64]. Thus she knows that *if* she forms an expectation, it will *then* be justified, because the expectation would (together with her desire for self-knowledge) justify *itself*:

> Although the agent's expectation of acting is a conclusion to which he jumps before the evidence is complete, he jumps with the assurance that the conclusion will achieve verity even as he lands. [Velleman 1989, p. 64]

Thus, an intention is a belief that is justified because (the agent knows) it combines with one's desire for self-understanding. And when the execution of the intention is successful, the contents of the belief will be true, so that the agent *knows* that he is doing φ.[174]

The problem with Velleman's claim to epistemic credibility is that it appears to achieve too much. If we accept that it is a good defense of the ability to form intentions without prior evidence, then it follows that it would equally be epistemically permissible to form some patently irrational, or even pathological, beliefs. Imagine an anxious student, with low self-esteem, studying for a test. There may be no evidence to support the idea that the student will fail the test. Yet on Velleman's view, it seems that it would be permissible for the student to form the belief that he will fail the test anyway. For she may know that, if she forms the belief that she will fail the test, her anxiety will become uncontrollable, and she will fail. So the agent may jump to the conclusion, 'I will fail the test', 'in the assurance that the conclusion will achieve verity even as he lands'. Yet this seems the paradigm of irrational belief precisely *because* the only reason the expectation of failure will come true is that very expectation. Velleman wonders why the 'rules of justification' [Velleman 1989, p. 63] should require that we not form beliefs without prior evidence:

> Why would rules designed to help one arrive at the truth forbid one to form a belief that would be true? What errors would one be avoiding by refusing to form a belief that wouldn't be erroneous? [Velleman 1989, p. 63]

[174] As Rödl notes, there is actually an important problem for Velleman's account here. For the contents of one's belief is that one *will* φ. And, 'from the fact that I, in whatever way, know I will do [φ], it does not follow that I know that I am doing it while I am doing it' [Rödl 2007, p. 61]. So Velleman does not show how an agent can answer the question 'Why are you (presently) doing φ?'. I will ignore this problem for the sake of argument, but will discuss the temporal structure of an agent's practical knowledge in §4.3.4.

Surely, the answer to the latter question is that one would be avoiding the error of forming a belief that has no rational basis in fact. To bring this out, consider that it may even be that the agent has all kinds of decisive reasons to believe she will pass the test—he has studied hard, is one of the top students in the class, and the teacher is known to be lenient in grading exams. Still, it may be true, and known to the agent, that *if* he forms the belief that he will fail, it will become true that he fails. Of course, once he *has* formed this belief, the circumstances have changed, and the evidence may henceforth favor the proposition that he will fail. But forming that belief in the first place would fly in the face of all the evidence, and *as such* would constitute an epistemic misstep.

So it seems that there is *prima facie* reason to doubt the epistemic credentials of the procedure of belief-formation that Velleman imagines. However, what if there is no prior evidence either way? It may be that there is simply no evidence either that our student will fail, or that she will pass. In that case, forming the belief that he will fail wouldn't fly in the face of the evidence. So wouldn't it be permissible? One problem with this response is that there are certainly cases in which one may intend to do something that the prior evidence does not predict. For example, an agent may know that she is not very punctual, and that it will be difficult to make her three o'clock appointment. Knowing herself, the agent may think, it doesn't make much sense to suppose she will be on time. Yet she may intend, even successfully, to make it to her appointment on time. So Velleman cannot use the defense that forming a self-fulfilling belief is permissible, when doing so doesn't fly in the face of prior evidence.

It seems that fundamentally, the problem for Velleman's account is that it does not properly distinguish intentions from mere predictions. That is why obviously irrational predictions, like our student's belief that he will fail, come out as epistemically credible: since it really would be rationally permissible to form an *intention* to fail, even if the prior evidence is that one will succeed, Velleman has no means to make the student's mere *belief* that he will fail come out as irrational, because an intention, for him, just *is* a self-fulfilling belief. Of course, intentions were supposed to be distinguished from mere self-fulfilling predictions by the fact that the expectation of self-fulfillment of an intention is grounded in the agent's knowledge of her desire for self-understanding—whereas for our student, this expectation is grounded in his knowledge that he has an anxiety disorder. But this cannot really account for the distinction. For, as Bratman has noted [Bratman 1991, pp. 261-262], if we have a desire for self-knowledge that combines with beliefs about what we will do to cause that very outcome, we may expect this desire to conspire with *any* belief about what we will do—including, for example, the belief that one will be late, will

fail the test, or will crash one's bike ('I'm going too fast!'). We could acquire more self-knowledge by making those beliefs come true. So any belief about what one will do—even those that are obviously not intentions—can combine with the desire for self-understanding, causing an action which then renders the belief true. And actions caused in this way are obviously not intentional.[175]

Now Velleman might insist that intentions are distinguished not just by the fact that they *do* interact with the desire for self-understanding to cause the action, but also that this fact is represented in the belief: intending to φ is believing that one will φ *because* of the desire for self-understanding. But given that we might expect *any* belief to interact with the desire for self-understanding to cause the action it represents, it seems that it would be perfectly reasonable for our student to expect that, if he forms the belief that he will fail the test, that belief will combine with the desire for self-understanding, causing him to fail (e.g., by making him restless and unable to concentrate).[176] He could then form the belief: 'I will fail the test because this very belief will interact with my desire for-self understanding'. Yet obviously, he would not *intend* to fail the test—he just expects things to unfold that way.[177]

4.2.2 Setiya: intentions as desire-like beliefs

Let's begin with a small recap of my criticism of the intention-as-belief approach thus far. The problems for Velleman's account seem to stem from the fact that the causal route from intention to action goes through the separate desire for self-understanding. Because of that, he cannot distinguish between what are plausibly intentions, and what are just (irrational) predictions of what will happen. *Any* beliefs about one's conduct can be turned into self-knowledge if the desire for self-understanding causes the action it represents. This 'problem of promiscuity', as Bratman calls it, [Bratman 1991, pp. 261-262] seems inescapable. For the reason that there is no way in which the desire for self-understanding can discriminate between intentions and other beliefs is that what is supposed to *make* this distinction, for Velleman, *is* precisely the causal

[175] This shows that the present problem for Velleman's account is a variant of a scenario of so-called causal deviance. I develop this kind of worry for the intention-as-belief view in more detail below (§4.2.3).

[176] Perhaps this problem could be surmounted by stipulating that the desire for self-understanding only causally interacts with beliefs of the form 'I will do φ because of my desire for self-understanding'. However, that would be rather *ad hoc*. Why would agents have a desire to know precisely those things that they will do because of that very desire?

[177] One might think that the distinction between mere beliefs and intentions is made by the fact that the latter include a reference to preexisting desires: the content of an intention is 'I'm hereby reinforcing those predispositions [to φ] to the point where I'll do it next', as Velleman [1989, p. 88] said. To see that this does not make a difference, just imagine that our student has a (perhaps not very strong) preexisting desire to fail the test so he can be rid of his anxiety: if he fails, he will not make it into law school, and will not have to take any further tests in the future. He may believe that this desire, combined with the belief that he will fail, will unnerve him so much that he will be unable to concentrate, and hence fail—without, of course, failing intentionally.

involvement of the desire for self-understanding. So perhaps the intention-as-belief account could do better with a different story, on which an intention does not need the help of a separate desire to cause what it represents. Setiya offers an account that promises to do precisely this. In this section, I will explain his account, delaying my argument against it until §4.2.3.

Setiya starts with the idea that there are two distinctive marks of intentional action. First, an intentional action is done for a reason, and thus gives application to a particular sense of the question 'Why?'. The second mark is that intentionally acting agents conform to the following requirement:

> *Belief:* When someone is acting intentionally, there must be something he is doing intentionally, not merely trying to do, in the belief that he is doing it.[178] [Setiya 2007, p. 26]

The challenge, for Setiya, is to explain how these two marks relate:

> . . . why should action done for reasons, or to which a certain sense of the question 'why?' is given application, necessarily satisfy *Belief*? What is it about being-done-for-reasons—or being susceptible to the question 'why?'—that requires the presence of belief?

Setiya supposes that there is something in the nature of a reason for φ'ing that entails that the agent must believe he is φ'ing (or at least, that there is *some* action he believes he is doing). An account of intentional action must explain this relation. As an Anscombean, I can only agree. Let us see how Setiya makes sense of the conceptual connection he identifies.

Setiya [2007, pp. 32-33] argues we can never account for the connection as long as we understand reasons for acting in terms only of pro-attitudes or desires combined with a means-end belief, as the standard version of CTA (see §1.2) does . For there seems to be no link at all between desiring that φ, and believing this can be achieved by doing ψ, that implies that one believes that one *is* doing either φ or ψ. So Setiya suggests the belief-desire view must be wrong, and understanding why it is wrong will help us to understand the connection between reasons for acting and *Belief*. In his diagnosis of the failure of the belief-desire view, he appeals to an example of Velleman's, which the latter takes from Freud:

> In one narrative, Freud's sister tells him that his attractive new desk is marred only by his old inkstand, which does not match. Later, Freud sweeps the inkstand on to the floor, with a peculiar and remarkable clumsiness.[179] On Freud's interpretation,

[178] Setiya [2009], in response to [Paul 2009], later weakens this principle to accommodate cases in which agents merely have an increased *confidence* that they are doing what they intend to do. This nuance need not bother us here, since my argument in §4.2.3 will apply even to the stronger version of Setiya's account.

[179] Setiya here cites [Velleman 2000, pp. 2-3].

> and Velleman's, Freud is moved to break the inkstand by the belief that breaking it will persuade his sister to buy him a new one, and the desire for a matching inkstand. [...] Despite all this, Freud did not act intentionally in breaking the inkstand. [...] Although it would be right to say that Freud knocked the inkstand with his arm out of a desire to break it, it would be wrong to say that he did so *in order to* break it, or on the *ground* that doing so would break it. [Setiya 2007, pp. 33-34]

According to Setiya, the problem is that the kind of explanation we give of Freud's action in this case is 'one of mere psychological motivation' [Setiya 2007, p. 34], and not one of 'acting on the basis of a reason' [Setiya 2007, p. 34]. That is, Anscombe's question 'Why?' just does not apply to it. Of course, this is a strong claim, and it seems that it stands in need of more argument, if Freud's case is to defeat the standard belief-desire model. I will attempt to reconstruct this argument below (§4.2.3). For now, let us grant Setiya that Freud's case is not one of acting intentionally, and that the belief-desire view thus does not capture what it is to act for a reason, over and above being psychologically motivated.

What, then, must be added to mere psychological motivation, in order to arrive at action on the basis of a reason? Setiya suggests—rightly, it seems—that in the case described above, what is missing is that Freud 'does not *see* the considerations by which he is moved *as* good reasons to act' [Setiya 2007, p. 35]. He explains:

> ... while Freud may believe that he has a justification for breaking the inkstand, this belief is not involved in the behavior by which he does it. When he knocks the inkstand to the floor, his desire is acting through him, without his warrant. [...] When I decide to go for a walk because the weather is fine, I *take* that fact about the weather *as* my reason to act. [Setiya 2007, p. 36]

Setiya's suggestion is that accounting for the phenomenon of 'taking as a reason' will explain why *Belief* is true. This is because the 'taking' in 'taking as a reason', on his view, has 'at once a *practical* meaning [...] and an epistemic one' [Setiya 2007, p. 39]:

> *Both* meanings are involved when we say that someone who acts because p takes p as his reason to act. In doing so, he thinks of p as his reason to act, and he is moved by this recognition. Taking something as my reason is a kind of 'desire-like belief'. It is a belief-like representation of p as my reason to act, and at the same time a decision to act on that reason... [Setiya 2007, p. 39]

A desire-like belief is thus a representation of the reason one has for doing φ that is causally efficacious in producing φ. Granting that such states are possible,[180] we

[180] A desire-like belief is something for which the simple belief-desire model has no room: beliefs and desires are supposed to be distinguished by their direction of fit, i.e., their different causal relations to the world. See [M. Smith 1987, p. 54] for an argument that beliefs and desires must be 'distinct existences'. However, Setiya [2007, pp. 49-50] argues that desire-like beliefs *are* possible. Compare Millikan's [1995]

can see how they might improve on the standard belief-desire model. Where Freud, in the original case, was moved by his belief that he could get a new inkstand by breaking the old one, he was not moved by a *recognition* that this was a reason for him to act. On Setiya's proposal, Freud could have a representation of himself *as* knocking down the inkstand *because* he wants a new one—which representation then goes on to cause the action of knocking down the inkstand. Setiya thus proposes the following, preliminary, account of taking something as one's reason to act:

> To take p as one's reason for doing φ is to have the desire-like belief that one is doing φ because of the *belief* that p—where this is the 'because' of motivation.[181] [Setiya 2007, p. 43]

The proposal is preliminary because having a desire-like belief with the above contents does not suffice for acting for a reason. Again, take Freud's case:

> . . . Freud takes it that he is breaking the inkstand because his sister will then be moved to buy him a new one, where this is the "because" of [psychological] motivation. In other words, he takes himself to be motivated as he was in the original version of the case. But then he need not take himself to be acting for a *reason*, in being moved by this belief. . . [Setiya 2007, p. 45]

On the preliminary proposal, the only difference with the original, non-intentional Freud-case, is that Freud now *knows* (well-trained in the art of psycho-analysis as he is) that he is being psychologically caused to break the inkstand by his desire for a new one. That is hardly the kind of attitude we have to our intentional actions. Thus, Setiya argues:

> If the content of taking-as-my-reason is to depict me as acting for a reason, not just as being motivated by a belief, it must depict me as being motivated by the way I *take* the consideration that p. [Setiya 2007, p. 45]

What must be altered in the preliminary proposal is, therefore, the *sense* of the 'because' in the agent's desire-like belief. The desire-like belief of the agent's must represent herself as being motivated *by* her taking-as-a-reason, and that is, *by* that very state itself. The contents of the state must thus be *self-referential*. Incorporating this insight, Setiya formulates his ultimate account of taking as a reason:

> To take p as one's reason for doing φ is to have the desire-like belief that one is *hereby* doing φ because of the *belief* that p. [Setiya 2007, p. 46]

The 'hereby' signals self-referentiality: the agent represents herself as doing φ because of the belief that p 'by way of being in this mental state' [Setiya 2007, p. 46].[182] If

so-called 'pushme-pullyou representations'.

[181] That is, it is the 'because' of 'mere psychological motivation' [Setiya 2007, p. 34].

[182] Compare Harman [1997].

Freud were to be in *that* mental state, and if that really were the cause of his breaking the inkstand, it seems he would really be motivated by his taking himself to have a reason. Setiya thus goes on to identify such a self-referential desire-like belief with the agent's intention in acting [Setiya 2007, p. 48]. *Belief* then follows straightforwardly: since an intention is a desire-like belief in which the agent represents herself *as doing* φ (for the reason that p), an intentionally acting agent will always have a belief about what she is doing intentionally.[183]

Moreover, Setiya argues elsewhere [Setiya 2008, p. 408], such a belief can be sufficiently warranted to count as knowledge. As we have seen in our discussion of Velleman (§4.2.1), his account attempted to defend the epistemic credentials of intentions by defending that it is *in general* permissible to form beliefs without prior evidence, as long as one knows that *forming* that belief will make it come true. Setiya argues for a weaker principle: in general, prior evidence for belief is required, but some beliefs—intentions—are exempt from this condition. One exempting condition is an agent's having *knowledge how*[184] to do what he intends [Setiya 2008, p. 408]. Knowledge-how performs the role that prior evidence performs for ordinary beliefs: it provides the warrant required to *form* a new belief. And once it is formed, the belief is of course justified because of the reliable connection between intending and doing: the intention represents the agent as being in a state that causes φ, so whenever the agent is actually in that state, he is (and knows that he is) in a state that tends to bring about φ. Whenever nothing goes wrong with the execution of the intention, the agent will thus have practical knowledge.

Before I move to criticize his account (§4.2.3), it is interesting to note that Setiya [2007, p. 24] claims that it shows that practical knowledge is 'spontaneous': it is knowledge which is not acquired by observation, but by acquiring an intention which causes what it represents. And Setiya suggests that this spontaneous character of intentions 'will be welcomed by the incompatibilist' about free will:

> [The incompatibilist] doubts that freedom can be reconciled with causal determination, but [. . .] wants to make sense of acting for a reason nonetheless. It is sometimes argued that this cannot be done: an agent's acts for reasons only [in] so far as they correspond to causes that determine his action. [. . .] It may be true that reasons correspond to determining causes, but they are causes that operate only when one intends them to—and one's intention need not be determined or even caused. This is not to say that the present account *depends* on indeterminism, or that it conflicts with a compatibilist conception of freedom. [Setiya 2007, p. 59]

[183] As Setiya notes, his account is easily transposed to the case of intention for the future, in which the agent does not think 'I hereby *am doing* φ because of the belief that p', but 'I hereby *will do* φ because of the belief that p' [Setiya 2007, p. 49].

[184] Setiya [2012] argues that know-how is irreducible to knowledge-that.

In the light of what we have seen in chapter 1, it seems like the incompatibilist (specifically, the libertarian) should *not* welcome Setiya's account. For as we have seen there, accepting an account of intentional action that is *compatible* with determinism, such as Setiya's, makes it impossible to insist that freedom requires indeterminism. And although I think Setiya is right that freedom consists in the spontaneity of practical knowledge, I also believe that what he calls spontaneity does not deserve the name. For whether or not one *forms* an intention, on his variant of CTA, is itself either explained by causes which operate behind the agent's back (in the case of determinism) or is fully accidental (in the case of indeterminism). As I will later argue (§4.3.2 and §5.3), the spontaneity of practical knowledge does not consist in being moved by causes which by which one intends to be moved, in the 'ordinary' sense of causation. Rather, it consists in a special *form* of causality: acquiring practical knowledge is an agent's determining *herself* to act.

4.2.3 Three levels of accidentality

In this section, I will argue that the intention-as-belief approach does not succeed in giving a sound account of practical knowledge, and *a fortiori*, of intentional action. The focus will be on Setiya's version of the approach, as we have already seen that it has certain advantages compared to Velleman's. However, the argument will not, ultimately, depend on the particularities of any specific version of the view. As we will see, there is a fundamental flaw in the idea that we can account for intentions as beliefs causing the action they represent. This flaw runs deep, and, as I will argue (§4.3), stems from the fundamental commitments of CTA itself.

First, let us return to Setiya's criticism of the standard belief-desire model of intentional action, and his diagnosis of the Freud-case:

> Freud is moved to break the inkstand by the belief that breaking it will persuade his sister to buy him a new one, and the desire for a matching inkstand. [. . .] Despite all this, Freud did not act intentionally in breaking the inkstand. [Setiya 2007, p. 33]

As I said, Setiya's claim that this case is a counterexample to the belief-desire model seems to require more argument: just *why* should we take the case as one in which the agent does not act intentionally? Now there is a plausible answer to this question. On the face of it, the problem for the belief-desire model presented by this case is an instance of a deviant causal chain.[185] *If* the case is not one of intentional action, despite all the conditions of the belief-desire view being met, it seems this must be because Freud's desire and means-ends belief do not cause his knocking down the inkstand 'in the right way'. But interestingly, Setiya denies that the case is one of

[185] See my introduction of the topic of deviance in §2.2.2.

causal deviance: 'this causation is non-deviant, since the desire to break the inkstand guides the movement of Freud's arm in sweeping it to the floor'[186] [Setiya 2007, p. 33]. Instead, Setiya goes on to argue that the problem is that Freud simply does not act for a reason:

> There seem to be two levels of explanation here: one of mere psychological motivation, which does apply to Freud, and one of acting on the basis of a reason, which does not. [Setiya 2007, p. 34]

This is surprising, because the problem posed by cases of deviant causation is precisely that the agent does not act on the basis of a reason, even though she is 'psychologically motivated' (i.e., she has a causally effective desire for the outcome). Of course, in famous cases of deviance, such as that of Davidson's climber (see §2.2.2), there is an intermediate state—the climber's nervousness—which combines with his desire to be rid of the weight in order to cause the outcome—the climber's letting go of the rope. In Freud's case, there is no such intermediate state. Still, it would be wrong to conclude that both cases do not expose the same flaw in the Davidsonian account. For both cases show that causation by a belief-desire pair is not sufficient to ensure that an action is *non-accidentally* related to the agent's reasons. In fact, we can view Setiya's proposed account of intention as belief as an attempt to eliminate this non-accidentality, just as Davidson's own account is an attempt to eliminate it at an earlier level. We can see this by briefly revisiting the original rationale behind CTA.

Davidson argues against the idea that action-explanation is non-causal along the following lines. Suppose that an agent has a reason to raise her arm—she desires to catch a friend's attention, say, and believes she can accomplish this by arm-signaling. Now just at that moment, a queer spasm causes her arm to go up, catching the friend's attention. Now the agent's desire 'rationalizes' the arm-rising event in some sense—the agent might think it good luck that that which she desired actually happened, when she happened to desire it. But of course the agent did not act intentionally: her desire and what happened just coincided accidentally. Not so, argues Davidson, if we suppose that the desire was the *cause* of the arm-rising: then the two are related non-accidentally, and the action is thus intentional. Causation is what is supposed to make the difference between φ'ing while having some reason to do so, and φ'ing *on the basis of* that reason.

Now as Davidson recognized, just adding a causal link between reason and action is actually *not* enough to account for an agent to be acting on the basis of a reason. This is shown by examples like his climber case. In such cases of deviant causation, the agent's reasons accidentally cause the very action they happen to rationalize. In

[186] For more on Setiya's idea that the causation is not deviant because the movement of Freud's arm is 'guided', see below.

order to eliminate this second level of accidentality, Davidson argues that what we must have is 'causation in the right way':

> Wanting to do something of type x may cause someone to do something of type x, and yet the causal chain may operate in such a manner that the act is not intentional. [...] Beliefs and desires that would rationalize an action if they caused it in the *right* way—through a course of practical reasoning, as we might try saying—may cause it in other ways. [Davidson 1973, p. 79]

As Davidson makes clear, the problem with the climber case is that the agent's reasons—his belief and desire—do not operate *as* his reasons. Although the climber does as he desires, and does so because he desires it, his desire does not play the role of a premise in the agent's practical reasoning. But what is it for a belief and a desire to cause an action 'through a course of practical reasoning'? Davidson does not know himself, but, we can now see, Setiya's proposal answers precisely that question. For what is missing in the climber case seems to be the same as what is missing from Freud's case: the causal chain from reason to action does not operate through the agent's *recognition* that he *has* a reason. In the climber case, it operates via a different mechanism, i.e., his nervousness. In Freud's case, there is no alternative intermediate state, but the causal chain is still not 'right', because it does not involve the agent's *taking* himself to have a reason. So we may view Setiya as having discovered (inspired by Anscombe, of course) that what it is for a reason to cause an action non-accidentally, through a course of practical reasoning, *is* for an agent to perform the action based on the recognition that she has reason to do it.

This should remind us of the kind of causality I discussed in the previous chapter: that of theoretical reasoning. It is no accident, I argued, that someone who believes that p because she believes that q represents herself as believing the conclusion on the grounds of the premise. Similarly for practical reasoning and intentional action: it is not an accident that an agent who φ's because of ψ represents herself as doing the former because of the latter. Furthermore, the causal connection she represents between her reason and her action must operate *through* the agent's representation of that very causality. For otherwise, the agent's representation would be only accidentality related to her action: she would represent herself as being moved by e.g. a desire without endorsing that desire as her *ground* for doing it. So an intentionally acting agent represents herself as acting because of a reason, where this 'because' is the because of practical reasoning, or action-explanation. We may thus formulate the following principle:

> (**Intentional Representation**) If A is φ'ing intentionally, then she represents herself as φ'ing intentionally.

I believe that this principle—(IR) for short—is true. The principle is stronger than

Belief, but of course, Setiya also endorses it: that is why he adds the self-referentiality condition to his preliminary proposal (see §4.2.2). Indeed, for Setiya, the truth of (IR) seems to be the *explanation* of *Belief*. The problem for his account of intention, I argue, is that it cannot really accommodate the truth of (IR). Setiya does not succeed in preventing a third iteration of the same problem of accidentality that threatened Davidson.

Consider the example I described in the discussion of Velleman (§4.2.1). A student studying for an exam believes that if she would believe that she were to fail, that belief would in fact cause her to fail. On Velleman's account, I argued, such a pathological belief would be epistemically warranted, since the agent might expect the belief to interact with the desire for self-knowledge to bring about the agent's failure on the test, and would have to count as an intention—although it obviously isn't. We can now see that the problem posed by the example is in fact a form of causal deviance: the agent's beliefs causes the event it represents, but not in the right way, that is, not through her recognition that she has *reason* to do it. So if an agent believes that she will φ because that very belief will combine with the desire for self-knowledge to cause her to φ, she does not satisfy (IR).

Now notice that Setiya's analysis of intention as a desire-like belief does not improve matters. Suppose that, if the agent comes to expect that she will fail, this thought will distract her so that she will fail the test. The agent may believe that this is the case: being aware of her own nervous character, she may believe that having the thought that she will fail will cause her to fail ('just as it happened last time', she may think). And suppose that even if it is not true that she is not good enough to pass the test, her *thinking* that she will fail because she is not good enough *will* cause her to fail. So she could have the following thought: 'I believe that I will hereby fail the test because I believe I'm not good enough'. This belief may distract her to such an extent that she fails the test. It is then a desire-like belief[187] with contents of the form that Setiya says constitute an intention: 'I am/will hereby do φ because of the belief that p'. But, it seems, the agent does not take her belief that she is not good enough as a *reason* for failing the test, nor does she take herself to act intentionally. In fact, it seems she may think: 'now that I've had this thought, it is inevitable that I will fail', while at the same time insisting that she does not *want* or *intend* to fail. Quite the opposite: she wants to pass, but her anxiety about not being good enough frustrates her efforts. It follows that having a desire-like belief that conforms to Setiya's account of intention does not satisfy (IR). An agent who has such a desire-like belief does not *thereby* represent herself as acting intentionally.

Let us consider some objections to this argument. First, one might object that our

[187] It is desire-like in the respect that it goes on to cause an action. See fn. 180.

example does not constitute a problem for Setiya's account of intention, because the student's thinking 'I'm not good enough' is obviously not the sort of thing to take the place of *p* in 'I'm hereby doing *φ* because of the belief that *p*'. For shouldn't the belief that *p* be something that somehow *rationalizes* the action? If so, there would be a relevant difference between intentions, as Setiya conceives of them, and the student's belief. But this objection is mistaken. Although it is of course *true* that believing 'I'm not good enough' is not really a reason for action (except perhaps in very special circumstances), Setiya cannot appeal to the distinction between what is a reason for action and what is not in this context—just as Anscombe could not appeal to the notion of a reason for action in order to illuminate the question 'Why?' (see §4.1.1). For his burden of proof is to explain what it is means for an agent to take something as a reason. The schema 'I'm hereby doing *φ* because of the belief that *p*' *is* his account of having a reason for action. That is why the example of the student provides a counterexample.

Second, one might object that whatever solutions to the problem of causal deviance are available to other variants of CTA, Setiya might make use of as well. True, the literature on causal deviance is huge, and many putative solutions have been put forward. Why could one of these not deal with our example? One problem with this reply is that it again pretends that it is one thing to solve the problem of causal deviance, and another to give an account of acting for a reason. As I explained in §2.2.2, this is false: to give an account of what it is for a reason to cause an action non-deviantly *is* to account for the non-accidentality that is intentional action.[188] And as I argued above, Setiya's account of intention should in fact *itself* be viewed as a way Zf eliminating deviance or accidentality from CTA. But perhaps, the objector might insist, we will arrive at a sound account of acting for a reason if we supplement Setiya's account of intention with extra conditions against deviance. Indeed, Setiya himself suggests the following way of dealing with e.g. Davidson's climber case:

> . . . the crucial concept is that of guidance: when an agent *φ*'s intentionally, he wants to *φ*, and this desire not only causes but continues to guide behaviour towards its object. Sustained causation of a process towards a goal is not unique to intentional action: it is present in purposive behaviour that is not intentional. So although it is something of which we lack an adequate theory, there is no circularity in taking it for granted here. [Setiya 2007, p. 32]

Yet it seems obvious that sustained causation will not solve the problem of deviance in the case of our student. For why not suppose that her desire-like belief sustains her inability to give the right answers to the test's questions for the entire time during which she takes the exam? It is of course true that intentions causally sustain the

[188] Also see Horst [2015].

actions they represent. But *since* that is not unique to intentional action, as Setiya admits, causal guidance cannot be the ingredient that is missing from the student's case. Now many other accounts of the 'right' kind of causal chain have been put forward in order to solve the problem of deviance. Unfortunately, there is no space to discuss them all.[189] However, in §4.3 I argue that there is fundamental reason to believe no solution is forthcoming.

Let us take stock. We have moved through several iterations of the problem of causal deviance. Each time, the problem is that the connection between the agent's desires or intention and her action is merely accidental. If there is no causal relation between the reason or intention and the action at all, it is a mere accident that they coincide. But a causal connection between reason and action may still be accidental, if the causation does not operate 'through a course of practical reasoning'. For the causal relation to be non-accidental, Setiya suggests, is for it to operate through the agent's *recognition* that she has a reason to act. That is, the non-accidentality characteristic of intentional action consists in the agent's acting on a *representation* of herself as acting on that very reason. Setiya tries to cash this out in terms of a desire-like belief with self-referential contents. What the example of our student shows is that this is not enough. Setiya's proposal just moves the accidentality up a level.[190] In our example, the agent's self-representation does not cause the action 'in the right way'. So, we may conclude, intentional action requires that *it is no accident* that the self-representation causes the action. In §4.3, I will argue that eliminating accidentality requires us to conceive of practical knowledge as an *unmediated* act of self-knowledge.

4.3 Self-knowledge of acting intentionally

I argued that (IR) was a necessary requirement of intentional action. For an agent to act on the basis of her reasons is for her to act on the *recognition* that her reasons speak in favor of performing the action. Thus, in acting on a reason, an agent is conscious of acting on a reason, and that is, of acting intentionally. As we have seen, Velleman's and Setiya's accounts of intention as a self-fulfilling belief are unable to accomodate this principle. In this section, I will argue that we can only accommodate (IR) if we follow Anscombe in thinking of practical knowledge as a form of unmediated self-

[189] For an excellent and comprehensive overview of proposed solutions to deviance, I once again refer to van Miltenburg [2015, pp. 91-130].

[190] I believe these repeated iterations of the problem of accidentality show that what I suggested in §2.2.2 is true: intentionality *is itself* a form of non-accidentality, and therefore it will not be possible to account for this non-accidentality in terms of ordinary causal relations that satisfy certain further requirements. The argument in §ch4-pk-self reinforces this claim.

knowledge. First (§4.3.1), I will provide a deeper diagnosis of why variants of CTA fail to account for (IR), and thus for intentional action. As we will see, the reason for this is that on the intention-as-belief view (and on CTA more broadly speaking), the relation an agent has to her φ'ing can only be *receptive*. I will then (§4.3.2) explain and argue for the alternative Anscombean way of understanding practical knowledge as productive.

4.3.1 Wittgensteinian arithmetic

Return to the case of our student, who thinks she is hereby failing the test because she thinks she's not good enough. I suggested that the problem here was that intentional action requires that it is no accident that the agent's self-representation causes the action. Let us try to develop this diagnosis in more detail. To do so, it will be instructive to inquire into the logical form of the contents of an intention, on the intention-as-belief view. We can represent this form quasi-schematically as follows:

> I will do/am doing φ, and φ is caused by mental state(s) $X, Y, \ldots$.

Velleman and Setiya each offer a different specification of this schema by giving a different account of the required mental states. But what their accounts have in common is that the variable φ does *not* stand for an intentional action. Rather, it stands for an *event* (a bodily movement) which is supposed to be intentional in virtue of the fact that the predicate 'caused by ___' applies to it. The first part of the thought, the clause 'I am doing φ', does not represent the agent as performing an intentional action. It would be question begging to say that it represents her as performing φ *un*intentionally, for on the intention-as-belief approach it is supposed to be an open question whether φ is intentional or not—that just depends on whether φ meets the relevant further (causal) conditions. So let us, borrowing a term from Anscombe [Anscombe 1963, p. 28],[191] say that φ stands for a *preintentional* event. So the logical form of the thought in our schema is this: it predicates the concept 'caused by ___' of the preintentional event φ. Notice that adherence to this schema is not optional for the intention-as-belief view. In fact, it follows from what Lavin [2013, p. 278] calls the *decompositional* structure of any variant of CTA. CTA, as we have seen 1.2, proceeds as if the problem of action theory is to give an answer to Wittgenstein's famous question 'What is left over if I subtract the fact that my arm goes up from the fact that I raise

[191] I take it that Anscombe's argument in that paragraph, to the effect that 'We do not add anything attaching to the action at the time it is done by describing it as intentional' [Anscombe 1963, p. 28], is trying to make the same point as my argument below: describing, or representing, an action as intentional is not representing it as a mere event of which we predicate certain further conditions.

my arm?' [Wittgenstein 2001 [1953], §621]. In Lavin's words, CTA is an exercise in 'Wittgensteinian arithmetic'[192] [Lavin 2013, p. 277]:

> It insists that the theory of action be a search for a solution to this equation: action [. . .] consists of a not-intrinsically-intentional physical event, a *mere happening*, occurring in a context where certain further facts obtain. [Lavin 2013, p. 277]

Thus, a decompositional analysis supposes precisely that intentionality must be an 'extra feature' [Anscombe 1963, p. 28] that is attached to a preintentional event—a thesis which Anscombe denies, and for good reason, as I will argue here.[193]

Now there is an interesting and unjustly neglected question about the kind of reference that thoughts of our schematic form are supposed to exhibit. For it seems that in order to predicate the concept 'caused by ___' of the event φ, the agent must *already* stand in a cognitive relationship to φ: as I explained in §3.4.1, an act of predication[194] requires a prior act of *reference* to the object in question. For example, in order to predicate 'is red' of a tomato in front of oneself, one must first be able to cognitively pick out the tomato, e.g. demonstratively, by standing in a perceptual relationship to it. This is characteristic of what we, following Rödl, called *receptive* reference. If the same goes in this case, it follows that the agent already cognitively singles out an object—the event of her φ'ing preintentionally—*before* judging that the predicate 'caused by ___' applies to it. And that is to say: she cognitively singles out her φ'ing *before judging that it is an intentional action*.

That seems to be precisely what is wrong in the case of our student. She somehow knows (or believes) she is performing a preintentional action falling under the description 'failing the test', and applies the concept 'caused by this mental state' to it. That he is φ'ing 'because of ___' is something he comes to *realizes* in an act of predication *separate* from his singling out φ as an action-description he satisfies. She must already be somehow aware of φ'ing (preintentionally), and needs to *identify* this as something he is doing intentionally (or not). *That*, to use Velleman's phrase, is the essence of his passivity with regards to his φ'ing.

Velleman and Setiya may claim that their account of intention-as-belief accommodates the insight that intentionally acting agents have knowledge of what they do without observation. But there is a real question of *how* the referential relationship to

[192] It is interesting to note that CTA is not the *only* theory of action that engages in such arithmetic. For example, Ginet's [2008] non-causal theory does the same: it says, approximately, that an intentional action is a mere movement occurring in the context where an intention is *present*—although it need not cause the action.

[193] Notice that the attempt to give such a decompositional analysis means that (in so far as the proponent of CTA has an analysis of practical knowledge at all) it is committed to the practical analogue of the assumption I criticized in chapter 3: the assumption that a subject's self-knowledge and the object of that self-knowledge are Distinct Existences (DE). More on this analogy in §4.3.3.

[194] Unless, of course, we are dealing with self-predication in an act of unmediated self-knowledge. See §4.3.2.

an agent's preintentional φ'ing comes about, if not through observation.[195] They do not seem to meditate on this question, because it is so rarely realized that the variable φ must stand for a preintentional event—arguably, because it is just dogmatically assumed that there cannot be a formal, intrinsic difference between intentional action and mere events. But in fact, as I will argue below, an agent's reference to her intentional action is an unmediated act of self-predication, wherefore an agent does not need to *settle* the question whether her φ'ing is an intentional action.

Of course, Velleman and Setiya *want* to agree with the idea that an intentionally acting agent does not need to settle the question whether her φ'ing is intentional or not. But it is not obvious that their accounts can deliver this. For it does not seem to matter *what* an agent predicates of the preintentional event of her φ'ing: the problem is that on their accounts, there *is* a gap between her realization that she is doing φ and her representing φ as intentional at all. This is why Anscombe, in criticizing Davidson, writes:

> He speaks of the possibility of 'wrong' or 'freak' causal connexions. I say that any recognizable causal connexions would be 'wrong', and that he can do no more than postulate a 'right' causal connexion in the happy security that none such can be found. If a causal connexion were found we could always still ask: 'But was the act done for the sake of the end and in view of the thing believed?' [Anscombe 1974, pp. 110-11]

Her confidence that there is *no* 'right' causal connection stems from the fact that, on CTA, a representation of an intentional action is a representation of a preintentional event to which certain further conditions apply. And as long as this is the case, *that* the agent predicates 'caused by ___' is *accidental* to her representation of φ. Such a predication never amounts to a judgment that the action is intentional, because it demonstrates the agent's need to *settle* the question whether her φ'ing is intentional or not. And of course, this is what generates the possibility of causal connections that are *obviously* wrong: in cases like Davidson's climber or our student, the agent's judgment that 'I'm φ'ing (preintentionally)' is *made true* by what happens—the unintentional loosening of the rope or failing of the test.

It will be interesting to compare this diagnosis of the problem of deviance with another one that has recently become popular in the literature. That is the idea that what is necessary to make the causal chain 'right' is 'causation in virtue of content'.[196] For example, here is Schlosser's diagnosis of the climber case:

[195] I am here assuming that the reference to the preintentional action is demonstrative. Perhaps there are alternatives: might the agent not refer through a definite description (e.g., 'the action I am hereby causing')? It is difficult to see how this would help. Reference through a definite description is still *receptive* reference: it is an attempt to cognitively single out an object out of a range of given objects. And as I argued in §3.4.1), receptive reference cannot underlie first person thought.

[196] Apart from Schlosser, who is cited below, proponents of this approach include Wedgwood [2006], Arpaly [2009].

> The reason why the causal pathway in the climber example is deviant seems to be that the movement is merely caused by the reason states—the movement is not a response to the reasons qua reasons. [The action is not] a response to the reason state qua reason state; it does not show that the reason state causes the action because it is a reason state. The reason state causes and rationalizes the action, but the reason state's rationalizing the action seems to be irrelevant to its causing it. [Schlosser 2007, p. 191]

This diagnosis seems perfectly correct. As I said, in the climber case, the agent's reasons cause the action—but accidentally, i.e., not through a course of practical reasoning. Setiya's account was supposed to remedy this: as such, it is an attempt to explain what it means to say that an action is a response to a reason state *qua* reason state. But we have seen that in the case of our student, for example, the intention still causes the failing of the test accidentally. In that case, too, it seems that the action is not a response to the intention *qua* intention. The intention causes and represents the action, but its representing the action seems to be *as* irrelevant to its causing it as the fact that the climber's desire rationalizes the action of loosening the rope is irrelevant to the fact that this desire (or its onset) causes the action. So I agree that intentional action requires causation in virtue of content. The question is whether CTA can deliver this. Consider again the example of our student.

Of course, it is true that the student's self-representation causes her failing the test. But does it cause the action *in virtue* of representing it? It seems not. The student represents himself as doing something which is caused by a self-fulfilling belief. Yet he does not represent the fact *that* the belief is self-fulfilling as obtaining *in virtue* of his representation of it *as* self-fulfilling. He thinks 'I will fail because I have this thought', and the thought goes on to cause his failing the test. But the fact *that* the belief goes on to have this effect is something that happens independently of the agent's representation. (Perhaps the effect is causally mediated by the student's nervousness, or perhaps the belief just has that effect, like Freud's desire for a new inkstand just had the effect of causing his arm to sweep the old one to the floor.) And this lack of causation in virtue of content, I argue, is something that cannot be remedied on CTA.

On CTA, content and causation—or equivalently, causation and representation—are strictly separated.[197] This is obvious from the very notion of a propositional attitude. *Qua* physical state, a propositional attitude stands in causal relations to other physical states. But *qua* mental state, it possesses *propositional content*. Now different attitudes can have the same propositional content. One can both *believe* and *desire*, e.g., 'that the door is open'. *What* a state represents as being the case is one

[197] These considerations are due to [Horst 2015].

question, and the direction of causation to or from the world ('direction of fit'[198], as it is often called) is quite another. So the propositional content a state bears is *accidentally related* to the causal role it plays: it is never *in virtue* of bearing a certain content that the state goes on to cause something. In fact, the whole point of the exercise, for Davidson, was to *explain* what it means for an agent to do or believe something *on the grounds of* something else—and that *just is* for something to be a response to a reason *qua* reason—to ordinary causal explanation. Causation in virtue of content cannot be *part* of an analysis of the 'right' causal chain, because it is in fact the very *explanandum* of CTA.

Compare this to Anscombe's point. Anscombe says that for any causal chain from reason to action, it would always be an open question whether the act was done 'for the sake of the end' [Anscombe 1974, pp. 110-11]. That is to say: it is always an open question whether the action was a response to the reason *qua* reason. I argued that the reason for her confidence in this claim is that a representation of a preintentional action *plus* certain further conditions never amounts to a judgment that the action is intentional, because it demonstrates the agent's need to settle the question whether or not it is intentional. So there is no such thing as a 'right' causal chain: as long as content and representation are separated, as they must be on CTA, it will be an accident that an action is caused by a state which also represents it.[199] Therefore CTA is mistaken: it does not give a correct account of the kind of non-accidentality manifested in intentional action.

If the above is correct, then representing oneself as acting intentionally is *not* judging that one is engaged in a preintentional event and then judging that certain further conditions apply to it. Rather, judging that one is φ'ing must *already* be judging that one is doing it intentionally. And, we can now see, giving an account of a kind of representation of one's own action that guarantees that it is intentional, and accounting for causation in virtue of content, are one and the same. For a representation of φ as intentional *is* a representation of it as a response to a reason, *qua* reason. In the next section, I will explain how we can make sense of such responsiveness to reasons *qua* reasons.

4.3.2 The cause of what it understands

In §4.3.1, I offered a diagnosis of the failure of the intention-as-belief approach: since, on CTA, representation and causation are conceived of as independent, an agent's representation of what she is doing can never amount to a representation of her action

[198] See §4.3.4 for more on direction of fit.

[199] Compare §2.2.2, where I first suggested that it is futile to try and define intentional action in terms of a non-deviant causal chain.

as a *non-accidental* response to her reasons. So no representation of the way that her (preintentional) φ'ing is caused will amount to a representation of φ as intentional. This means that the connection between the agent's reasons and her actions will *always* be accidental—not just in cases of obviously deviant chains. So on a sound account of intentional action, it must be no accident that an agent's representation of her φ'ing causes that action. It is not enough if representation and causation merely coincide, as they do on CTA. Rather, representation and causation must be (re)united: the agent's representing the action must *be* her causing it. Only then will the relation between the two be non-accidental. In such a representation that one is doing φ, φ will be represented as intentional *directly*—not in virtue of any further conditions being true of it.[200]. In this section, I argue that there is indeed such a form of representation: practical knowledge is an act of unmediated self-predication.[201]

In the previous chapter, I argued that a theoretical inference is an essentially self-conscious form of causality. I suggest that *practical* self-knowledge must be understood analogously. It is no accident that someone who reasons, 'p, so q' has knowledge of grounds that she believes q because she believes that p. Rather, her knowledge of grounds *is* her making the inference—judging that q follows from p is representing oneself as believing the one on the grounds of the other. We arrived at this conclusion by reflection on Boyle's idea that an inference doesn't deposit beliefs in one's mind like a stone. Rather, an inference continues to support belief in the conclusion as long as the subject subscribes to it. And it does so *through* the subject's knowledge of grounds: knowing that she believes that q because she believes that p is 'an enduring actualization of [the subject's] capacity to hold a proposition true for a reason she deems adequate' [Boyle 2011a, p. 22]. So a theoretical inference is an example of the kind of representation we are looking for: a representation that is a form of causality. As I said, a subject's reasoning 'q because p' and her knowledge of grounds are not two acts of thought, but one: her self-applying the concept 'believing that p' *is* her judging that p is right to believe, *is* her concluding that p is true. And this kind of representation defines the very notion of belief: something is not a belief that one does not represent as part of one's total view of reality. Hence, theoretical self-knowledge is the *form* of belief: something is not a belief that the subject does not represent *as* a belief.

To see that this also applies to practical self-knowledge, consider again the idea that an agent φ's intentionally if she acts on a *recognition* that she has reason to φ. For example, she wants or intends to ψ, thinks 'I should φ because I want to ψ', and concludes by doing φ. It may seem that there are *three* terms that need to be connected

[200] Compare Anscombe [1963, §19].

[201] The argument in this section borrows extensively from Rödl [2007, pp. 56-57], to whom I am, again, greatly indebted.

here: an intention, a normative thought, and a concluding action. The first two, then, need to be causally connected to third. But that is precisely the mistake that leads to the failure of CTA. If 'I should do φ because ___' is one of the *causes* of the action, then it must cause it non-accidentally, or 'in the right way'. The only substance that can be given to 'the right way' is: through a course of practical reasoning. And that is, it must cause it through the agent's recognition *that* the state provides her with reason to φ—thus leading to a regress. So the normative thought 'I should do φ because ___' must not be a third *term* in the reasoning at all. The agent's recognition that she has reason to φ because she intends to ψ is not a further *cause* of her action. Rather, it is the *connection* between her intending to ψ and her φ'ing. Thinking the normative thought and doing φ are *a single actualization of the agent's capacity to reason practically*. An action is an agent's answer to the question what to do, and her determining what to do is her representing it as to be done, just as her determining what to believe is her representing it as right to believe.

So φ'ing for the reason that (I want to) ψ *is* thinking 'I should do φ because I intend to ψ'. It follows that practical reasoning is a source of practical knowledge. For in thinking 'I should do φ because ___', the agent represents φ as her answer to the question what to do. Knowing that φ is the conclusion of her practical reasoning is knowing that she is φ'ing. Thus it is no accident that a practically reasoning agent has knowledge of grounds: her thinking 'I intend to ψ, so I should φ' *is* her thinking 'I am doing φ because I intend to ψ'. The truth of the latter thought, the action-explanation, is determined by the subject's accepting the normative thought.[202] As in the theoretical case, there is no separating the causality and the representation of the causality. That is why Anscombe says that practical reasoning and the question 'Why?' reveal the same order: the 'because' in an answer to the question 'Why?'—'I'm doing A because I'm doing B'—*is* the 'therefore' in 'I'm doing B, therefore I should do A'.[203] Reasoning practically from ψ to φ, the agent represents *both* as her answer to the question what to do, just as in reasoning theoretically from p to q, an agent represents both as her answer to the question what to believe. Practical reasoning is deriving one action from another. That is why, for example, Thompson [2008, pp. 85-146] argues that the fundamental form of action-explanation is 'I'm doing φ because I am doing ψ'.

Thus *practical* reasoning, too, is an essentially self-conscious form of causality. That is what it means to say that representation and causation are one. If this is so,

[202] In §3.4.2, I asked the question: how can we say that the subject thinks *both* 'it is right to believe p because of q?' and 'I believe p because of q', while at the same time saying that these thoughts are identical? The answer there was that these two phrases explicate each other's form of predication. The same goes in the practical case: 'I ought to φ because I intend to ψ' explicates the form of predication in a subject's practical thought 'I am doing φ because I am doing ψ'.

[203] See §4.1.2.

then an agent's thinking 'I'm doing *A*', when this expresses her practical knowledge, is an act of unmediated self-predication. For it is true *that* the agent is doing φ *because* she knows this practically. The knowledge is not a reflection of an independent reality: her φ'ing is not a state of affairs that obtains *prior* to the agent's knowing it. Indeed, that is what Anscombe means when she says that practical knowledge is 'the cause of what it understands':

> Practical knowledge is 'the cause of what it understands', unlike 'speculative' knowledge, which 'is derived from the objects known'. This means more than that practical knowledge is observed to be a necessary condition of the production of various results [...]. It means that without it what happens does not come under the description—execution of intentions—whose characteristic we have been investigating. [Anscombe 1963, pp. 87-88]

Practical knowledge is the form of action, just as theoretical self-knowledge is the form of belief.[204] Therefore practical knowledge that one is φ'ing is not receptive: the state of affairs does not need to affect the subject in order for her to know. The agent comes to know that she is doing φ by making up her mind about what to do. So the agent's knowledge does not involve an act of reference separate from an act of predication: she does not need to *identify* the person she judges to be doing φ. Instead, her judging that she is doing φ is an act of self-predicating the concept 'doing φ'. She equally does not need to identify a (preintentional) event of φ'ing in order to predicate of it that it is caused in such-and-such a way. Self-predicating the concept of 'doing φ' *already is* representing φ as an intentional action: just as someone who makes a Fregean judgment '*Fa*' *thereby* thinks 'the object *a* falls under the concept *F*', so an agent who self-predicates 'doing φ' thereby represents herself as acting intentionally. An agent does not need to *settle* the question whether φ is an intentional action, because her representation of φ is not accidentally related to her practical reasoning.

Notice that the idea that practical knowledge is an act of unmediated self-predication is unavailable to adherents of the intention-as-belief approach. For that approach is wedded to the decompositional picture of intentional action as a preintentional event to which certain further conditions apply. That means that *whether* an agent is doing *X* must be true *independently* of whether she represents herself as φ'ing.[205] The representation and the action must be an independent reality. Therefore any knowledge the agent can have of what she is doing must be receptive.

[204] It thus seems that there is a sense in which Anscombe's opposition between practical and speculative knowledge is misleading. For that theoretical self-knowledge is the form of belief means that it is, in some sense, also the 'cause of what it understands'—what it understands is just not an action. The fundamental opposition is thus between receptive knowledge and unmediated self-knowledge, of which self-knowledge of belief and practical knowledge are two species.

[205] I further argue for this claim below (§4.3.3).

At this point, it is interesting to recall Anscombe's introduction of the topic of practical knowledge:

> Can it be that there is something modern philosophy has blankly misunderstood: namely, what ancient and medieval philosophers meant by *practical knowledge*? Certainly in modern philosophy we have an incorrigibly contemplative conception of knowledge. Knowledge must be something that is judged as such by being in accordance with the facts. The facts, reality, are prior, and dictate what is to be said, if it is knowledge. [Anscombe 1963, p. 57]

It seems that to suggest that an intention is a *belief* that one is doing φ is to commit precisely the mistake that Anscombe is here warning us about. It is to suppose that when it comes to the question whether an agent is or is not doing φ, '[t]he facts, reality, are prior, and dictate what is to be said'. Ironically, the intention-as-belief approach—for all its attempts to accommodate such a thing as productive knowledge—thus appears to be a prime example of the modern 'incorrigibly contemplative conception of knowledge'. Letting go of that conception in the philosophy of action requires that we abandon the decompositional orthodoxy. I will further explain how to do so in §4.3.3.

4.3.3 Action concepts and the concept of action

An intentional action is an agent's response to her reasons *qua* reasons. So, I argued, action-explanation and its corollary, practical reasoning, must signify a form of causality that *is* the agent's representation of her reasons—the kind of causality inherent in unmediated self-knowledge. From that it followed that practical knowledge cannot be a mental state of belief, as beliefs are conceived on CTA and orthodox philosophy of mind. For a belief-state can never be unmediated self-knowledge. After all, a belief-state is a reality independent from what it represents. In this way, our reflections on the nature of an agent's representation of her intentional action, and the non-accidentality such a representation requires, have lead us to the conclusion that we must reject the most important dogma of contemporary philosophy of action: that intentional action can be decomposed into a mere, non-intrinsically intentional happening and a number of further conditions. Adapting Anscombe's phrase, an intentional action is not a fact prior to the agent's knowledge of it. Rather, it is a non-accidental *unity* of thought and movement. An intentional action is, literally, a '*thought that is a movement*' [Rödl 2007, p. 19].

Note again that the decompositional orthodoxy is in fact the analogue in the practical domain of the thesis I have criticized in the last chapter: the Distinct Existences (DE) assumption. That assumption was that believing something and knowing one-

self to believe it are two distinct acts of mind. Similarly, a decompositional analysis of intentional action has it that doing something and knowing oneself to be doing it are two distinct acts. Now we have seen in §4.1.2 that there is a worry about rejecting (DE) in the practical case: an action is so obviously a *physical*, external world *happening*—a replenishing of the water supply, or a sinking of the Bismarck. How could that be the same reality as an agent's knowledge of it?

We should recognize that this worry begs the question against the Anscombean account I have been developing. It simply presupposes the idea that a real-world happening cannot be a manifestation of an agent's capacity for practical self-knowledge. Yet that is precisely what Anscombe says *is* possible. As Thompson says:[206] [Thompson 2011, p. 200]

> The overarching thesis of *Intention* was that self-knowledge [. . .] extends beyond the inner recesses of the mind, beyond the narrowly psychical, and into the things that I am doing.

Although the worry may be question-begging, it is still important to address it. Our argument in §4.3.2 may show *that* decompositionality is false, but we want to understand what this means, and how it squares with the intuition that action is physical happening. Anscombe addresses this matter as follows:

> What bearing can what the agent thinks have on the true description of what he does? Someone may want to say: if what he does is a happening, a physical event, something 'in the external world', then that happening must be something that takes place, whatever the agent thinks. If you give a description of it, for the truth of which it matters what the agent thinks, such as 'He got married', 'He swore an oath', 'He murdered his father', then your description ought to be broken down into descriptions of thoughts and of purely physical happenings. If we ask: Why? the answer is: because what an agent thinks simply cannot make any difference to the truth of a description of a physical fact or event. But, if you say this, it only shows what you mean by 'a physical fact or event'. [Anscombe 1991b, p. 4]

We may, of course, *define* 'physical happening' so that it trivially follows that no physical event is the same reality as an agent's thought about it (i.e., her practical knowledge). But then this doesn't show anything about the cases Anscombe takes as an example: intentional acts like marrying, promising, and murder. They might still be happenings that are *not* independent of the agent's thought. And indeed, it is easy to see that for these examples, the decompositional strategy Anscombe is discussing—breaking them down into descriptions of thoughts and of purely physical happenings—fails.

[206] Thompson's quote explains why Anscombe's slogan, 'I *do* what *happens*' [Anscombe 1963, p. 52], is so important to her. Doing something intentionally, or intending to do it, is knowing oneself to be doing it—and knowing it *is* doing it.

Consider the example of marriage. Anscombe [1991, 1969] suggests that all the truth-conditions for, e.g., the action-description 'she is getting married' may be in place except that the bride does not know that she is doing so (she thinks she is only going through a rehearsal). Then no marriage is in fact taking place. The 'purely physical' descriptions are all there, but what is missing is the agent's thought. But precisely what is it that the agent must think in order for the sum of the purely physical happenings and her thought to add up to a marriage? It cannot just be that the bride must think that all the purely physical truth-conditions (that she is in a church, etc.) are satisfied: she might think *that* even in the scenario where she takes herself to just be going through a rehearsal. Rather, her knowledge of what she is doing must include the fact that *she is getting married*. But then the action-concept 'getting married', which she applies to herself, cannot be the concept of a 'purely physical' event: *what* she knows herself to be doing, when she knows she is getting married, must *include the fact that she knows she* is *getting married*—for that is evidently one of the truth-conditions of getting married.

The decompositional strategy thus fails for cases like marrying, promising, and murder. What is special about such actions is that they are examples of things that one can *only* do intentionally. Anscombe [1963, p. 85] gives other examples of action-types for which this holds: e.g., 'greeting', 'paying', 'hiring', 'sending for'. She contrasts these with action-types like 'sliding on ice', or 'offending', which one can do *either* intentionally or unintentionally. So one might think that, while the decompositional strategy fails for action-types which can only be performed intentionally, it may yet succeed for the latter category. This is mistaken. To see this, consider a typical thing one might do intentionally or unintentionally: flipping a switch.

It may seem obvious that for this kind of action, what *happens* is the same, whether or not it happens intentionally. If an agent flips the switch intentionally, there is the *same* 'purely physical' event (a human hand presses against a button on the wall) that might occur unintentionally, *plus* the agent's thought. And in this case, unlike the marriage example, it seems it is not difficult to say what it is that the agent must think: the action-concept she must (self-)apply is the concept of a purely physical happening, namely, a switch-flipping. But in fact this is mistaken. Suppose the agent is thinking, e.g., 'It is the case that I flip the switch'. Now she has a queer spasm, causing her arm to hit the wall and flip the switch. The agent satisfies the 'purely physical' description 'flipping the switch', and she self-applies the concept of flipping the switch, but she does not flip the switch intentionally. What we have here, of course, is another case of deviance or accidentality. As we have seen (§§4.2.3 and 4.3.1), it is hopeless to try and keep adding 'purely physical' descriptions of what happens to the agent's thought. What the agent must think, in order to flip the switch

intentionally, is precisely *that she is flipping it intentionally*.

Thus, the same holds for both the category of things one can only do intentionally, and the category of things one can do intentionally and unintentionally. *If* φ is an intentional action, then the concept of φ'ing that the agent self-applies must *include* the fact that she knows she is φ'ing. Not only that: it must also be part of the concept that she satisfies the concept *because* she knows she satisfies it. This formally distinguishes a concept of an intentional switch-flipping from the concept of a mere event that we may describe using the same form of words. Call the former type of concept an *action concept*. The agent's intentional activity is *part* of an action concept: part of the concept one self-applies in acting intentionally *is* that one is acting intentionally. Therefore, we might say that an action concept includes the concept of action, i.e., it includes the concept of a movement that is constituted by an agent's practical knowledge.

The distinction between action concepts and concepts of mere happenings is obscured if we use the same form of words, 'doing φ', to stand for both kinds of concepts. That is the source of the pressure to think that the (DE) assumption must be true in the practical domain, even if it isn't in the theoretical case: in the case of action there *is* such a thing as flipping the switch *un*intentionally, while there is nothing analogous in the case of belief—there is no such thing as 'merely' being in a state that p, as opposed to believing it. There is therefore such a thing as representing oneself, first-personally, as flipping the switch unintentionally, and we may be mislead into thinking that such a representation is what must be added to a 'purely physical' happening to turn it into an intentional action. But as I have argued, this is a mistake. The intuition that actions are physical movements thus does not support (DE): rather, we must accept that some physical movements are identical to the agent's knowledge of them.

4.3.4 Knowledge of action-in-progress

There is another intuition that militates in favor of the (DE) assumption in the practical domain. As we discussed in §4.1.2, there seem to be cases in which an agent fails to execute her intention. If an intention is an agent's representation of what she is doing, it seems that in such cases, her representation is *false*. But that means the representation must be a reality independent of the action: the intention is perhaps aimed at correctly representing what the agent is doing, but fails. Moran has sharply formulated the worry:

> The agent can, of course, be flatly wrong about what he takes himself to be doing, for instance [...] when the pen runs out of ink without his noticing, or more generally

> when the empirical conditions enabling a particular action fail to obtain. If he is wrong in assuming that writing is getting produced, then he cannot have practical knowledge that he is writing. [Moran 2004, p. 60]

So it seems an intention must be a kind of *belief*—sometimes true, sometimes false—after all. How can the Anscombean account escape from this apparently simple argument? Consider Anscombe's own take on cases of failed action:

> I wrote 'I am a fool' on the blackboard with my eyes shut. Now when I said what I wrote, ought I to have said: this is what I am writing, if my intention is getting executed; instead of simply: this is what I am writing? [Anscombe 1963, p. 82]

Her answer is negative. Even if something had gone wrong with the chalk, and the intention thus failed to get executed, Anscombe insists 'my knowledge would have been the same' [Anscombe 1963, p. 82]. She justifies her claim that her 'knowledge would have been the same' by saying that, when an intention fails to get executed, 'the mistake is in the performance':

> . . . I say to myself 'Now I press Button A'—pressing Button B—a thing which can certainly happen. [. . .] And here, to use Theophrastus' expression again,[207] the mistake is not one of judgement but of performance. That is, we do *not* say: What you *said* was a mistake, because it was supposed to describe what you did and did not describe it, but: What you *did* was a mistake, because it was not in accordance with what you said. [Anscombe 1963, p. 57]

Clearly then, Anscombe is suggesting that, when the chalk breaks and no writing appears, the agent's judgment 'I am writing' is not to be faulted. This is commonly seen as an appeal to the special *direction of fit* of an intention. Indeed, Anscombe's [1963, pp. 56-57] famous example of the distinction between a shopping list—which is not mistaken when a grocery-shopper comes home with margarine instead of butter—and the list of a detective who writes down everything the grocery-shopper buys—which *would* be mistaken if it mentioned buttered rather than margarine—is often taken to be the *locus classicus* of the very idea of direction of fit. But the supposed appeal to direction of fit is also seen as a mistaken way of dealing with the problem at hand: that an agent might *think* she is, e.g., writing on the blackboard, while in fact she is not. For example, here is Moran:

> . . . it is *not* a good answer to this problem of error and knowledge to advert to Theophrastus and the thought that the mistake here lies in the *performance* and not in what is *said*. To disqualify as knowledge, it doesn't matter where the error comes from so long as there is error; 'direction of fit' considerations are not to the point here. [Moran 2004, p. 61]

[207] The first reference to Theophrastus' remark is made in [Anscombe 1963, p. 5].

For Moran, direction of fit merely determines what needs to be altered when thought and the world fail to correspond. On this interpretation, an agent's practical representation of what she is doing differs from ordinary belief only in that it is the world (one's behaviour) that must be adapted, instead of the belief.[208] But the representation, which stays in place while the subject goes on to try to get the world to correspond to it, is still *false*, and therefore not knowledge. Moran is here projecting the concept of direction of fit, as it appears in contemporary philosophy of mind and language, on Anscombe's doctrine that 'the mistake is in the performance'. And if Anscombe's appeal was indeed to *that* notion of direction of fit, Moran would be right that it fails to solve the problem of error and practical knowledge. However, we should remember the locution 'direction of fit' does not actually occur in *Intention*. What Anscombe seems to have in mind is different.

As we have seen, Anscombe rejects the idea that for practical knowledge, '[t]he facts, reality, are prior, and dictate what is to be said, if it is knowledge' [Anscombe 1963, p. 57]. Practical knowledge, for her, is the *same* reality as the action. This means more than that an intention has a 'world-to-mind' direction of fit. It is not that the agent's (false) practical representation will endure while she is busy getting the world to cooperate. Rather, it is the knowledge that *determines* what is to be said—i.e., under what description the agent's movement falls. As we have seen (§4.3.3), that is a feature of action concepts: the agent's knowing that she is φ'ing is part of the truth-conditions of the description 'she is doing φ' (if this is a description of an intentional action).

The idea that the mistake is in the performance when an action fails is therefore best seen as a rejection of the very notion of a *propositional attitude*. In contemporary philosophy of mind, e.g. believing and desiring are two distinct attitudes one might take to one and the same proposition.[209] The content of the judgment, the proposition, is what it is regardless of the attitude a subject takes to it. Not so on the Anscombean picture I have tried to develop in this chapter and the last: on this picture, a belief that p and intention to φ are each characterized by a different form of (self-)predication: believing something is representing it *as* right to believe (i.e., as *true*), and intending to do something is representing it is right to do.[210] That followed from the need to

[208] Moran is followed by e.g. Grünbaum [2009, pp. 14-15] and Haddock: 'The moral of the story is that if one fails to have knowledge because one's belief is false, the way to ensure that one has practical knowledge is to modify one's behaviour accordingly; there is no need to modify one's belief' [Haddock 2010, p. 325].

[209] We have already seen this in §4.3.1.

[210] And that is, representing it as *good*, in the sense I explained in the discussion of practical reasoning (§4.3). An Anscombean account of action thus implies the famous thesis that acting intentionally is acting *sub specie boni*. As we can now see, this is a thesis about the *form* of practical judgment, and not a naively optimistic thought about human psychology. See Boyle and Lavin [2010] for an insightful defense of this 'guise of the good' thesis.

reunite content and representation. The difference between belief and intention is thus not in the attitude taken to some content, but in the logical form of that content itself. That is why Anscombe insisted that desire is not part of the content of a premise in practical reasoning: that doing φ is something the agent desires is the form of her judgment 'I'm doing φ'.[211]

We can thus see that it is not the contemporary notion of direction of fit that Anscombe has in mind when she rejects the idea that failure of execution of an intention threatens practical knowledge. When an agent intends to write on the blackboard, while unaware that the chalk has broken, her representation is *not* false: it could not be, because what she represents herself as doing determines under what description her movement falls. But how can we say this? Is it not obviously wrong to say that the agent writes on the board?

The solution lies in recognizing that '[a] man can *be doing* something which he nevertheless does not *do*' [Anscombe 1963, p. 39]. That is, intentional action is subject to a certain distinction in *aspect*: one can either *be doing* a certain thing (progressive aspect) or *have done it* (perfective aspect). *Having cooked* a risotto, it follows that Alice *was cooking* a risotto in the past. But if Alice *is cooking* right now, it does not follow that she will have cooked a risotto later: she might change her mind and make some pasta instead, or she may simply be prevented from finishing the dish by an unfortunate accident (she is struck by lightning just when she was about to fry the onions). Anscombe's examples of answers to the question 'Why?' often have progressive aspect—'I am doing A because I am doing B'. Thompson suggests that the ongoing character of an agent's present intentional doings is essential to a correct understanding of practical knowledge:

> [Practical knowledge's] character as knowledge is not affected when the hydrogen bomb goes off and most of what the agent is doing never gets done. [. . .] My so-called knowledge of my intentional action in truth exists only and precisely when there is no [completed] action, but only something I am doing. [Thompson 2011, p. 209]

Thus, when one is writing on the blackboard with a broken piece of chalk, one's practical knowledge determines that one falls under the progressive description 'writing on the blackboard'. Such statements are not falsified 'if philosophers perchance arrange that the H-bomb goes off just now' [Thompson 2011, p. 209] and the intended sentence thus never gets *written*. The action of which the agent has practical knowledge is ongoing: she might, presently, look up to the blackboard, notice that something is wrong, and set out to find a better piece of chalk. If this were to happen, she would

[211] As I explained in §4.1.1, the notion of desire at play here is obviously not that of a craving for something sweet, but of a kind of desire that is informed by reasoning about what is right to do. Compare Anscombe [1993].

all the while be engaged in one and the same intentional action. Thompson points out that this possibility of going on and correcting is essential to our ability to act. He takes the example of an agent in the business of making ten carbon copies—an example Davidson famously used to dismiss the idea that intentional action requires even *belief* that one is doing something. According to Davidson [1978, p. 129], the carbon-copier intends to make ten copies in pressing down his pen on the pile of papers, but has no inkling whether he is actually making ten. But Thompson notes that an action of making ten carbon copies is ordinarily very different:

> [...] you write on the top sheet, trying to make a good impression to get through all the carbon, then look to see if your impression made it through all of them. If it did, you stop. If it didn't, you remove the last properly impressed sheet and begin again. If necessary, you repeat. Even the man who has to go through five stages is all along, from the first feeble impression, making ten copies of the document, and he knows it, all along. [...] The one who doesn't know it, Davidson's man, must be under some strange mafia threat: he gets one chance, no checking, and he's dead if he doesn't manage it. [Thompson 2011, p. 210]

Attending to the fact that 'the content of practical knowledge is something present, and thus something of which more is to come, perhaps including several attempts at it' [Thompson 2011, p. 210] will therefore allow us to see how an agent may know what she is doing, even if *at present* she has not successfully executed her intention. An agent may discover that the water supply has not yet been replenished because there was a hole somewhere in the pipe, and thereupon set out to fix the hole and pump some more water, knowing all the while what she is doing.[212]

It is important to note that the progressive/perfective distinction is, of course, not unique to intentional action.[213] We can, for example, say 'that something was falling over but did not fall (because something stopped it)' [Anscombe 1963, p. 39]. The omnipresence of progressive judgments ('that stone *is rolling*', 'this tree *is falling*')

[212] Setiya presents Thompson's contrast between the two carbon-copiers as follows:

> Thompson contrasts two carbon-copiers, one of whom will check and confirm that the copies are made, the other of whom has only one shot. The first carbon-copier knows that he is making ten copies, even if he does not know that the copies are going through the first time. The second carbon-copier may succeed in making the copies, but he is not doing so intentionally, since he can only succeed by luck. [2012, p. 301]

This is misleading. It is not that the first copier knows because he will check, and that the second does not know because he will not check. For the first copier *already* knows, even if lightning strikes before he is able to check. The second copier is a man in a putative Davidsonian world, containing only *events* to which the contrast between progressive/perfective is not even applicable. I take it that Thompson's point is that intentional action would be impossible in such a world.

[213] For this reason, Schwenkler's argument that an appeal to the progressive is not a good answer to the problem of error in practical thought because it threatens the common-sense idea "that human actions are events every bit as 'worldly' as the rolling of a stone and the fall of a sparrow" [Schwenkler 2011, p. 137] is puzzling. The *rolling* of a stone is not less 'worldly' for being an incomplete, ongoing process.

should urge us to take seriously the thought that ongoing *processes* are part of the furniture of the world.[214] Now it may seem strange to say that something that is essentially incomplete—a crossing of the street, say—is still *real*. For as long as there is still cross*ing* of the street, there is no particular event of street-crossing (by *X*, at *t*). How can something that has precisely *not yet* come to pass be part of one's *ontology*—one's list of what there *is*? But taking the progressive seriously means precisely that we recognize that there is a sense in which what is or was merely happening enjoys a kind of presence in what *has* already happened.

Is it not altogether too fanciful to suppose that it is *true* that an agent *is writing* when he is holding something manifestly unfit to produce letters on a piece of paper, or *is replenishing* when there is a hole in the pipe? This question only makes sense when we still presuppose that we can say what it is that an agent is doing independently of her practical thought—that is, if we are still in the grips of the decompositional dogma. And in fact, the insight that acting intentionally is an ongoing process provides us with another reason to reject that dogma

To see this, consider a completed event, such as the making of a risotto. On the decompositional view, this event is an intentional action if and only if it was caused by the agent's representation of that event—e.g., by her intention to make a risotto. But we can now see that if this was an intentional action, then the agent was acting intentionally throughout the time it took her to make the risotto—for instance, she was acting intentionally from the moment she started cutting the onions, to the point when all the onions were cut. Suppose the agent were struck by lightning at that point. Then no risotto would have been made, and thus, no event would have taken place that corresponds to what she represented in her intention ('to make a risotto'). In order to render the cutting of the onions intentional, on the decompositional view, we thus have to postulate a second intention: one of onion cutting. That is of course not problematic. But notice that we can repeat the story: what if the agent were struck by lightning just as she was moving her knife through the air? That event would not be represented by her intention to cut an onion, so we have to postulate an intention to move the knife through the air, just so and so.

The problem this poses for the decompositional view is the following. Take an arbitrarily small and precise stretch of a movement—the exact trajectory of the cook's knife from *a* to *b* during a very small interval. In order for the agent to count as moving the knife from *a* to *c*, she must have an intention to move the knife from *a* to *b* in precisely this-and-that way, in order to move it *c*. And this seems implausible:

[214] No doubt an attempt can be made to reduce all sentences in the progressive to sentences mentioning only complete events and to insist that aspect is merely a linguistic artifact. See Rödl [2012, ch. 5] for an argument, based on Kant's and Aristotle's accounts of time and change, to the effect that any such attempted reduction is bound to fail.

in performing any action, one is often oblivious to the precise movements one is making. Trying to turn on the light, for example, one gropes here and there on the wall in search of the switch. Although every bit of the movement one's hand makes is intentional, it is wrong to say that one had an intention to move one's hand precisely along the trajectory that it did move.[215] On the Anscombean account, by contrast, we can say that the event of the cook's hand-movement was intentional under the description 'cutting the onion', because that is what she was *already* doing. The agent's practical knowledge does not need to represent a completed event, but represents an ongoing process. This is arguably not possible on the decompositional view (at least not on CTA), because then the representation needs to be in place *prior* to the occurrence of the action it represents. On the Anscombean account, this is not the case. Self-applying the action concept 'doing φ' *is* being engaged in φ'ing. Thus the agent is already doing what she intends to do—cutting the onion—even if she is now, say, only moving the knife through the air.

4.4 Conclusion

In this chapter, I have argued for an Anscombean theory of action that radically differs from CTA. On Anscombe's view, intentional action cannot be reduced to a mere event in addition to ordinary causal explanation. Rather, there is a teleological structure *in* intentional action: the *A*-*D* order revealed by the question 'Why?', which is the same order as the one inherent in practical reasoning. This order is the object of an agent's practical knowledge.

I have looked at attempts to accommodate the phenomenon of practical knowledge within CTA: the intention-as-belief approach (§4.2). As we have seen, this approach fails, because it cannot satisfy the demand (IR): an agent must represent herself as acting *intentionally*. On Setiya and Velleman's view, this is not the case: a belief that one is doing something, and that this is (event-)caused by some state, does not amount to a representation that one is doing it for a reason. Intentional action requires that an agent's self-representation causes the action *non-accidentally*,

[215] Since an agent can, e.g., be 'flipping the switch' for the entire time that it takes her to find and flip it, I agree with Lavin [2013] that there is no such thing as what he calls 'teleologically basic action'—a first intentional stretch of an action, *X*, such that the agent does nothing with the intention to do *X*. Lavin argues that CTA is committed to such basic action because of its need to isolate a 'non-intrinsically-intentional' description of an action. He argues that isolating the basic action is in fact impossible because actions are ongoing processes, to which the progressive/perfective distinction is applicable. Although his argument is similar to mine, I think it is mistaken in a crucial respect. CTA is not committed to the existence of teleologically basic actions because it needs to find a non-intrinsically intentional event. For if CTA is true, *every* event is non-intrinsically-intentional. It is rather that the decompositional analysis ensures that one cannot deal with the progressive nature of practical representations, on pain of having to postulate intentions for implausibly small and precise movements, and CTA is hence forced to resort to basic actions.

and this will not be the case as long as the representation is a mere *belief*—for then the agent's relation to the action will always be *receptive* (§4.3.1). This means that CTA is inherently flawed: the 'right way' for a reason to cause an action is through a course of practical reasoning, i.e., through a representation of that reason *as* a reason. But such a representation is impossible on CTA. Its decompositional structure ensures that causation and representation are strictly separated, and so, that the action is never a cause *in virtue* of *what* the agent represents—the *content* of her reason.

In §4.3.2, I argued that reuniting representation and content means that practical knowledge is an act of unmediated self-knowledge. Like its theoretical sibling, practical knowledge is knowledge acquired by making up one's mind: in this case, about what to do. An intentional action *is* the agent's answer to that question—it *is* the conclusion of her practical reasoning. Therefore, I argued, an action is not a reality independent of the agent's representation of it. That means we must resist the decompositional analysis of intentional action.

In §4.3.3, I argued that the urge to insist on decompositionality—which is the practical equivalent of the (DE) assumption about theoretical self-knowledge—is rooted in the misleading fact that there is such a thing as, e.g., flipping the switch *un*intentionally, where there is no analogy of this in the case of belief. I argued that we must nevertheless sharply distinguish *action concepts* from concepts of mere happenings. The former *contain* the agent's activity: in self-applying an action-concept, an agent thinks of herself *as acting*—that is, in having practical knowledge, she knows *that* she has practical knowledge. In §4.3.4, I explained how Anscombe's theory can overcome the problem posed by the possibility of failure to execute one's intention. *Pace* Velleman's claim that Anscombe must be a reliabilist (see §4.1.2), the reason why such cases do no threaten practical knowledge is the *ongoing*, or *progressive* nature of intentional action. Combining these insights, we may conclude that an action-concept is the concept of an ongoing movement to which the agent's knowledge is internal—the agent's knowledge is the *form* of what happens.

In §4.1, we saw that Anscombe's inquiry is an attempt to illuminate the *sense* of the special question 'Why?', as it applies to intentional action. Anscombe attempts to explicate that sense by showing what it means for the question to have application. We can now see that the question 'Why?', when it is a demand for a reason for action, is applicable *only when*, and *in virtue* of the fact that, an agent *knows* it to be applicable. This means that practical knowledge is a distinctly un-Cartesian variety of self-knowledge: it 'extends beyond the inner recesses of the mind, beyond the narrowly psychical, and into the things that I am doing' [Thompson 2011, p. 200]. As I mentioned in the previous chapter (§3.5), there is an important sense in which unmediated self-knowledge is *spontaneous*: such knowledge is not a re-

flection of a prior reality, but springs from the subject *herself*, i.e., from her own consciousness. Anscombe's account thus allows us to see how spontaneity may extend 'into the things that I am doing'. But does the fact that an intentional movement is spontaneous—that it is the manifestation of a rational, self-conscious capacity—mean that it is also undetermined, in the sense at stake in the free will debate? In the next chapter, I argue that it does.

* * *

Chapter 5

Freedom and self-movement

The explanation of action differs fundamentally from what Davidson calls ordinary causal explanation, so I argued in the previous chapter. The latter pertains to a kind of causality that is blind to reason, in the sense that it functions independently of an agent's representation of herself as having a reason to do something. By contrast, I argued, reasoning in general, and practical reasoning specifically, is an essentially self-conscious form of explanation: an agent's reasoning 'I'm doing ψ, so I should φ' *constitutes* the truth of an explanation 'I am doing φ because I am doing ψ'. As I explained, this means that the intentional action is *spontaneous*. In this chapter I will finally argue that from the spontaneous character of intentional action, it follows that such action is free and physically undetermined. In acting intentionally, an agent is genuinely self-determining. And in contrast to other libertarians, I will argue that we even possess knowledge that we in fact do exercise such self-determination.

However, this ambition may seem puzzling at first sight. For on the Anscombean account I am defending, it seems there is a sharp distinction between reasons and causes. And many theories of action which have insisted on such a sharp distinction have been compatibilist. Therefore I start (§5.1) with a review of such non-causal theories of action. I argue that their commitment to compatibilism is made possible by what I call the *two-domains picture*, on which there is a division of labor between reasons and causes. I explain how Anscombe's account differs from this picture: for Anscombe, an agent's reasons are irreducible to ordinary, mechanical causes, yet they still play a *productive* role in the coming about of her action. This combination of the irreducibility and productivity of reasons, I argue, commits us to indeterminism about the workings of nature on levels of description lower than that of action-explanation. But that does not yet prove that intentional action must

be undetermined. For as I explain, even if there is lower-level indeterminism, happenings described on the higher level of action-explanation may still be perfectly deterministic. To prove the thesis that intentional action is undetermined, we must also rule out such higher-level determinism.

Before I can do so, however, I will first have to rebut a worry about the possibility of an Anscombean account of action. For the idea that reasons are not reducible to, e.g., states in the brain, while still being productive causes, may seem puzzling. Would that not somehow imply that reasons-explanations break the laws of nature? I argue that it does not (§5.2). I do so by defending a *dispositional* or *powers-based* account of causality and laws of nature. I introduce this theory of causation by means of Anscombe's [1971] influential essay on the topic. As I argue, the question how higher-level phenomena like an agent's reasons for action can produce physical events (like intentional action) loses its bite if we think of causality as the activity of substances. For then acting might be a *higher-level power* of certain substances: human beings. The existence of higher-level powers does not break the laws of nature or commit us to dualism, as long as there is lower-level indeterminism.

The powers-based account of causality will also provide us with a sound understanding of determinism and indeterminism (§5.2.4). As I argue, we must distinguish between powers whose manifestation can be *triggered*, and powers that do no have such a trigger. Powers of the latter kind are indeterministic powers. In order to defend the thesis that acting intentionally requires indeterminism, I must argue that the power to act intentionally is such an indeterministic power.

I start my argument for this claim in §5.3. I critically examine Steward's recent defense of the idea that the power to act must be indeterministic because it is a power for *self-movement* [Steward 2012a]. Although I am sympathetic to this idea, I argue that Steward cannot provide us with a strong enough notion of self-movement to support the requirement of indeterminism. I show that a stronger notion of self-movement can be found in the teleological powers of living organism (§5.3.1). I argue that such powers exhibit a strong form of non-accidentality, so that we can reasonably say that, e.g., animals move according to *their own laws*. Such self-movement is a form of *spontaneity*. However, I also argue that this form of spontaneity is not yet strong enough to defeat a certain sort of compatibilism—namely, the sort which claims that the power to act is a higher-level deterministic power.

In §5.4, I argue that it is different with the power to act intentionally. That power exhibits a yet stronger form of non-accidentality and spontaneity: the self-conscious causality of practical reasoning. As I argue, that form of spontaneity indeed implies that the power to act is indeterministic, and truly deserves the name of freedom.

5.1 Non-causal action theory and indeterminism

I argued that it was necessary to find an alternative to CTA in order to arrive at a satisfactory notion of intentional action as *free* action, and as requiring indeterminism. But, some readers may wonder, how can the Anscombean alternative be of any help to the libertarian? Traditionally, philosophers who have insisted on a strict separation between reasons and causes have seen no trouble in adopting a form of compatibilism about freedom. Many of those whose publications belonged to what Davidson called the 'strong neo-Wittgensteinian current of small red books' [Davidson 2001b, p. 261]—and Anscombe is often listed among those authors—espouse such a compatibilism. The choice to defend an Anscombean theory of action may thus seem ill-founded, from the perspective of the libertarian theory I am trying to develop. However, Anscombe's account differs significantly from that of her neo-Wittgensteinian colleagues (and others who hold that action-explanation is not ordinary causal explanation), making it an especially good starting point for a libertarian account of freedom. To do so, it will be helpful to understand how the strict separation between reasons and causes may seem to support compatibilism.

5.1.1 The two-domains picture

The appeal of the combination of non-causal action theory and compatibilism is *prima facie* understandable: to the extent that reasons are not causes, it seems, whatever *does* cause one's actions cannot undermine its intentionality, or one's agency over it. So if the action is caused deterministically, that need not threaten its freedom. A prime example of this kind of reasoning can be found in A.I. Melden [1961] (who was also one of Davidson's [1963] principal targets):

> Where we are concerned with causal explanations, with events of which the happenings in question are effects in accordance with some law of causality, to that extent we are not concerned with human actions at all but, at best, with bodily movements or happenings; and 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. The reason is simple, namely, the radically different logical characteristics of the two bodies of discourse we employ in these distinct cases—the different concepts which are applicable to these different orders of inquiry. [Melden 1961, p. 184]

Melden thus insists that the question what caused an action—e.g., raising one's arm—is mistaken. To be sure, we can ask for and give the causes of a bodily movement—one's arm going up. But if *C* is the cause of one's arm-rising, it does not follow that it is the cause of one's arm-raising. The concept of a cause is 'wholly irrelevant

to the understanding we seek' when we give an action-explanation. Cause and bodily movement, and reason and action, are concepts that belong to two 'radically different' bodies of discourse. Although he shies away from an explicit endorsement of compatibilism, the implication is clear enough:

> . . . if the argument is correct, determinism [. . .] is not false but radically confused.[216] So it is with indeterminism and libertarianism [. . .] The trouble in all these cases is that the applicability in principle of the causal model is taken for granted. [Melden 1961, p. 202]

The very idea that an *action*, as opposed to a bodily movement, could be causally determined or undetermined, is incoherent. But this means that the bodily movement might very well be determined:

> Suppose, however, it were possible in principle to predict with perfect accuracy, how could one then maintain that the agent could do anything other than what he does, that his behaviour is not subject to causal factors in precisely the same way in which this is admittedly true of the motions of some heavenly body [. . .]? [. . .] Certainly, if one knew the state of the nervous system and musculature, then one could predict, given such-and-such stimulation, that, say the arm would rise in the air. [Melden 1961, p. 209]

Although this is not an outright endorsement of the thesis of universal determinism, Melden seems to say that a bodily movement (the rising of one's arm, not the raising of it) might, at least sometimes, be determined by prior states of affairs. Human freedom consists in the fact that such a mechanistic account of what happens would say nothing about what actions we perform—not in the fact that we could not give a true account of that kind.

The picture, then, is one of two logical domains: a space of reasons and a realm of law, in McDowell's [1994] terms.[217] On this picture, the concepts of reason, action and freedom are protected against reduction to talk of mere events and blind causes by isolating them (at least conceptually) from the natural, or 'merely physical' world. Freedom is a concept at home only in the space of reasons. Perhaps 'free action' is synonymous with 'intentional action', similar to what I argue—or perhaps only a subclass of all intentional actions is free or voluntary. Either way, determinism is only a feature of the realm of law, and as such irrelevant to the question whether something is a free act.

[216] Melden seems to mean that the concept of determinism, and consequently that of indeterminism, is confused *as applied to human action*—i.e., he means that compatibilism and incompatibilism are confused. He further seems to use 'libertarianism' as a synonym for 'agent causation' [Melden 1961, p. 202].

[217] This is not say that McDowell himself endorses the kind of view I am here ascribing to, e.g., Melden and Von Wright. I think McDowell's ambition is precisely to overcome the duality between the law-bound and the rational or conceptual. See the introduction (p. 4) for discussion of how McDowell's project relates to mine.

The two-domains picture is illustrated by Stoutland:

> Determinism and freedom are compatible because they belong to [two] distinct types of explanation and hence two distinct ways of understanding and being. Our autonomy as rational agents is not compromised by any causal explanation of the effects of our acting.[218] [Stoutland 2009, p. 61]

Stoutland attributes this form of compatibilism to Von Wright, who was another prominent member of the little red book club Davidson argued against. Von Wright believes that an intentional action is necessitated by an agent's reasons: the premises of a practical syllogism logically *compel* the subsequent occurrence of the action.[219] Although the compulsion is not causal—since rationalization is a fundamentally different kind of explanation—this view is obviously amenable to compatibilism. The relation of logical compulsion between an agent's reasons and her arm-raising may exist in parallel to the causal compulsion of her arm-rising by, e.g., certain prior brain-events. For there is a division of labor between action-explanation and causal explanation: the latter explains the *occurrence* of some event, while the former aims to interpret or give meaning to an action.

For Von Wright, Melden, and other neo-Wittgensteinian compatibilists,[220] the very distinction between ordinary causal explanation and reasons-explanation that I have argued for is thus enough to ensure the irrelevance of the existence of a deterministic causal explanation of an event to the question whether that event is (under some different description) a free act. Part of the reason why this rejection of the relevance of determinism comes so naturally to these authors is that, for them, necessitation is precisely one of the features that distinguishes causal explanation from reasons-explanation. What it is for an event to have a causal explanation, on this view, is that it instantiates a universal generalization of the (approximate) form 'Always when *C* occurs, *E* follows'. That it is impossible to come up with such universal generalizations connecting reasons and action was thought to be one of

[218] Some may find that this way of representing the thought has a distinctly Kantian ring to it. And indeed, Stoutland [2009, p. 61] claims that Von Wright's compatibilism is Kantian. I unfortunately cannot go into the question to what extent Von Wright's account of freedom truly resembles Kant's.

[219] See Von Wright [1972], and Anscombe [1974] for a critique of Von Wright's account.

[220] Apart from Melden and Von Wright, we can also count Kenny [1976, p. 153] among these. Interestingly, Malcolm [1968] rejects the compatibility of what he calls 'mechanism'—which he defines as 'a special application of physical determinism' [Malcolm 1968, p. 45]—with reasons-explanation, and thus seems to lean to incompatibilism. Davidson [1963] also mentions some authors less directly associated with Wittgenstein as examples of the non-causal consensus he is arguing against, of which the most prominent are perhaps Hampshire and Ryle. Hampshire's treatment of action is to some extent sympathetic to the incompatibilist, arguing that rational reflection involves a kind of indeterminacy. However, in the end he seems to settle for a form of compatibilism, concluding that the indeterminacy of thought is only a feature of the reflexive standpoint we take up in deliberation [Hampshire 1975, pp. 140-141]. Ryle's views on determinism and compatibilism are unfortunately not very clear [Ryle 1949/2009, pp. 61-68].

the arguments in favor of the idea that action-explanation is not causal.[221,222] At the same time, it becomes natural to think that every event that *can* be causally explained (e.g., one's bodily movements) has a cause that, together with some universal law, necessitates the effect.[223] As the possibility of indeterministic causation shows, it is mistaken to assimilate causal explanation to necessitation in this way. We will return to this point below (§5.2.1).

However, the challenge for the libertarian remains even if it is allowed that ordinary causal explanation may be indeterministic: if reasons-explanation is not causal explanation anyway, what difference can the nature of the cause of a bodily movement, determined or undetermined, have to the intentionality and freedom of the action?[224] Of course, a non-causalist might still insist on indeterminism because he finds some of the standard arguments against compatibilism, such as the Consequence Argument, intuitively plausible.[225] But as I have argued (§1.4.1), appeals to incompatibilist intuitions about 'up-to-usness' and the value of indeterminism are ultimately unconvincing.[226] I argued that such an incompatibilist account of

[221] For example, see Kenny [1976, pp. 112-117] and Melden [1961, pp. 13-15]. The other famous 'little red book' argument against the identification of reasons with causes is the so-called 'logical connection argument', which holds that no *non-trivial* law connecting reason and action can be found, since if one φ's for reason X, it logically follows that the agent had reason X. For some versions of this argument, see Melden [1961, pp. 51-53] and Kenny [1976, pp. 117-120]. Of course, Davidson's anomalous monism (see §1.2) was designed to overcome both these problems. For our purposes, it is interesting to note that Anscombe never employs either of these classical arguments for non-causalism, making it unclear to what extent she belongs to the same school of thought as Melden, Kenny and Von Wright.

[222] Although Von Wright does think that the occurrence of an action can be logically deduced from the premises, he admits that the deduction can only be valid if we exclude such things as the action's being prevented, or the agent's changing her mind [Von Wright 1972, pp. 46-49]. The practical syllogism requires a *ceteris paribus* clause. So we cannot use a practical inference to predict with certainty that an agent will do X: we do not know whether other things will indeed be equal. However, reconstructing a practical syllogism may allow us to understand the point of someone's action after the fact: the syllogism then provides a 'schema of interpretation' [Von Wright 1972, p. 51]. The role of a premise in a practical syllogism, for Von Wright, is thus quite different from that of the antecedent of a universal law (which *can* paradigmatically be used to derive predictions). Of course, as is now widely recognized, the same difficulty of *ceteris paribus* clauses also plagues the idea that causal explanation is subsumption under universal laws. See e.g. Cartwright [1983]. Von Wright does not seem to have regarded this as a major problem for the Humean conception of causal explanation. In chapter 1 of his [1971], he seems to take that conception more or less for granted for the natural sciences, arguing only that understanding in the social sciences is of a different sort.

[223] Note that there is a tension between the view that causality is just the instantiation of a regularity and the idea that causes necessitate their effects. Indeed, the regularity-view is a neo-Humean way of accounting for such necessitation in ways that do not require modal connections, which the neo-Humean regards as suspect. More on this in §5.2.1.

[224] Indeed, some more contemporary defenders of a non-causal theory of action, whom we might expect no longer to be wedded to the idea that causality requires necessitation, are equally attracted to compatibilism. See for example Sehon [2012].

[225] A contemporary example of a non-causalist about action-explanation who is nevertheless an incompatibilist is for example Ginet [1983].

[226] Moreover, a non-causalist compatibilist seems easily able to refute the Consequence Argument. As I explained in §1.2.1, that argument attempts to show that if determinism is true, nobody had a choice about whether to φ. But if choice is understood along the non-causalist lines, as a non-reducible teleological notion, that inference becomes problematic. See e.g. Sehon [2012, p. 363].

intentional agency is unavailable as long as we stick to CTA. Unfortunately, it now seems that it is equally unavailable when we insist on the irreducibility of teleological explanation to ordinary causal explanation.

I suggest that this apparent dilemma is false. An incompatibilist must, as I have argued, reject CTA. That is not a sufficient condition for arriving at an incompatibilist understanding of agency, but it is a *necessary* condition. Accepting that reasons-explanation is irreducible to ordinary causal explanation, we can still insist that certain kinds of causal explanations (namely, deterministic ones) cannot apply to the events or bodily movements that are (under different descriptions) our intentional actions. That is, defending the irreducibility of reasons-explanation does not have to mean that we embrace the picture of two logically separate domains—a space of reasons and a realm of law that are fully independent of each other. To suppose otherwise is to tacitly adopt the assumption I criticized in §2.2.1: that there is but one form of causality.

In §5.1.2. I will argue that the Anscombean account of action I developed in chapter 4 allows us to reject compatibilism. For there is a fundamental difference between that account and the two-domains picture. As we have seen, an intentional action, on Anscombe's view, is a certain *unity* of thought and movement: there is no distinguishing between 'what merely happens' and the agent's practical knowledge of what she is doing. So if the agent has practical knowledge that she is doing φ because she is doing ψ, this reasons-explanation *constitutes* the movement she is performing. There is no opposition between real-world events, subject to the laws of physics, and actions, subject to the requirements of reasons. Rather, some real-world events are acts of (practical) reasoning, and as such are explained by the agent's reasons. From this perspective, the two-domains picture is thus *as* mistaken as the decompositional orthodoxy underlying CTA. For the two-domains picture actually *agrees* with CTA that a reason, *qua* reason, could not explain the *occurrence* of a real-world event.

It arguably follows that the two-domains picture suffers from a variety of the problem of accidentality which we have seen in §4.2.3. To separate the explanation of events, in terms of causes, and that of actions, in terms of reasons, is to make the agent's reasons *accidental* to the occurrence of the action. Put differently, on the two-domains picture, there simply can be no such thing as causation in virtue of content. And rejecting the idea of causation of virtue of content makes it difficult to make sense of the trivial point that actions *are* physical happenings, and that whatever is the explanation of an action must thus be the explanation of a physical happening. As I explained in §2.2.1, that was a true insight of Davidson's: a reasons-explanation must make what the agent does no accident. Thus the reasons cannot be idle in the

production of an action.

The idea that reasons are impotent to explain the occurrence of physical happenings events is alien to Anscombe. And that, I suggest, is why her account forms a fruitful basis for a libertarian theory of free will.[227] A short, preliminary look at her take on the issue of determinism and its compatibility with intentional agency (in §5.1.2) will thus help to set the stage for the more in-depth defense of incompatibilism I will undertake in the rest of this chapter.

5.1.2 Anscombe on the causation of action

As I explained, the ambition of Anscombe's account is to defend the irreducibility of reasons-explanation while rejecting the two-domains picture: she opposes the division of labor between reasons-explanation and ordinary causal explanation, whereby reasons do not explain the *happening* of actions, but only interpret them or place them in context. This raises a number of questions, most significantly how reasons can be irreducible to (ordinary) causes, while still having some kind of efficacy in the physical world. Anscombe's own writings on this subject [Anscombe 1983] will serve as our point of departure for answering this question.

Anscombe starts out with a seemingly traditional non-causalist insistence on the difference between two kinds of inquiry—action-explanation, and the explanation of mere bodily movement:

> When we consider 'the causation of action' we need to decide which sort of enquiry we are engaged in. Is it the physiological investigation of voluntary movements? I.e. do we want to know how the human mechanism works when, at a signal, the hand pushes a pen, or perhaps a door shut? [. . .] But that will not be our enquiry into the causation of action where our interests are in the following sort of question: What led to Jones' shutting the door then? We ascertain that he shut the door in order to have a private conversation with N. . . [Anscombe 1983, p. 102]

However, Anscombe's insistence that physiological investigation and explanations in terms of reasons should be strictly separated does not put her in the same camp as Melden and Von Wright. For while the latter sort of inquiry is irreducible to the former, Anscombe still thinks that there is a real question about how the two can coexist. She makes this clear by considering the point that the sorts of things one might find in the course of a physiological investigation—e.g. that an impulse in some

[227] That Anscombe herself was committed to incompatibilism is obvious from a number of writings. For example: 'My actions are mostly physical movements; if these physical movements are physically predetermined by processes which I do not control, then my freedom is perfectly illusory. The truth of physical indeterminism is thus indispensable if we are to make anything of the claim to freedom.' [Anscombe 1971, p. 146] This short passage already makes clear that Anscombe does not fit the mold of the two-domains picture.

nerves caused such-and-such a muscle to contract—may, under certain assumptions, themselves seem to lack real causal import:

> If we now think in terms of, say, some sort of elementary particles and the operation of the fundamental forces recognized by physics, the very descriptions which occur in physiology may seem to be descriptions of shadows. [Anscombe 1983, p. 102]

That is, supposing that the operations of nerves and muscles can be fully understood in terms of the operations of their constitutive parts (the ultimate 'atoms' of physics), it seems that descriptions of operations on the physiological level have at most an instrumental use: the *real* causal efficacy occurs at the lower level of description. Anscombe offers an example:

> . . . we really do find it scientifically convenient to speak of the causal efficacy of waves; they not only move but 'interfere' with each other. All the same, everyone will admit that this is just a convenient manner of speaking . . . the causal efficacy belongs rather to the masses of water particles . . . [Anscombe 1983, p. 103]

Just as it is a real question whether talk of causal interaction on the physiological level is more than a convenient manner of speaking, so Anscombe suggests there is a parallel question about reasons-explanation:

> . . . are we to consider the causality of action, when we are talking about histories of human dealings, as just a highly convenient, nay indispensable, *façon de parler*, such as we use also when we speak of waves as interfering with one another? I shall call descriptions in terms which in this way amount to a convenient *façon de parler*: *supervenient descriptions*. [Anscombe 1983, p. 103]

The fact that Anscombe stops to ask this question is important, for it shows that she is not committed to the two-domains picture. The description of events *as* intentional actions is *not* isolated from the workings of nature on a lower level. If descriptions of intentional actions *were* merely supervenient (in Anscombe's sense), then explanations in terms of reasons would be merely convenient ways of approximating the *real* explanations at lower levels of description.[228] It is true that reasons-explanation is different from explanation at lower levels of description in that it is teleological, i.e. it offers an interpretation or rationalization of someone's behavior. But (*pace* the

[228] It is important to note that Anscombe is here using the term supervenience in an idiosyncratic sense: she means to say that if determinism were true, everything that happens on higher levels of description would be explainable in terms of (and arguably, entailed by) states of affairs on the lower level of description. This is not to deny that there may be other forms of supervenience which would not make higher-level descriptions and explanations mere *façons de parler*. For example, so-called 'weak supervenience' is the thesis that there can be no difference in the higher-level properties of a thing without a difference in the underlying, lower-level state of affairs (the supervenience base). Compare Brian McLaughlin and Bennett [2014, §4.1]. There is no conflict between the possibility of intentional action and weak supervenience, nor does weak supervenience imply or depend on determinism. In what follows, I will use 'supervenience' in Anscombe's sense.

two-domains picture), talk of motives and intentions would still be no more than a *façon de parler* if causal efficacy belonged only to things on the level of description of physiology or physics.

Anscombe's point can also be understood in the following way. If the fact that an agent had a reason or intention to φ would not *make a difference* to her behavior, it would not really be an explanation of her action at all.[229] To adopt an example from Malcolm, suppose that an agent is climbing a ladder in order to retrieve his hat from the roof. If the man would have moved up the ladder *anyway*, even if he hadn't had the intention, then it seems that to give a explanation of his action in terms of his intention would be merely a *façon de parler*, in Anscombe's sense.[230] As Malcolm says, if 'the movements of the man on the ladder would be *completely* accounted for in terms of electrical, chemical, and mechanical processes in his body', then this would quite simply imply that 'his desire or intention to retrieve his hat had nothing to do with his movement up the ladder' [Malcolm 1968, p. 53].

So Anscombe is surely correct to say that reasons-explanation would be a mere *façon de parler* if an agent's behavior would supervene (in her sense) on the operations of, say, particles at the physical level—assuming, of course, that intentions cannot be *reduced* to those lower-level operations as on CTA. Anscombe calls the idea that reasons-explanation (and other purported forms of higher level explanation and organization) *is* a mere manner of speaking 'mechanism', which she characterizes as follows:

> there is some basic level of physical description such that all 'higher-level' descriptions are supervenient. Let the basic level be that of particles and fundamental physical forces. Then the forms of substances and animals and all sorts of actions and happenings will [...] be comparable to shadows. [Anscombe 1983, p. 104]

Now crucially, mechanism is *not* the same as determinism.[231] That is, the combination of mechanism and indeterminism is logically possible. The workings of the fundamental particles might be stochastic, while everything on higher levels of de-

[229] Anscombe says that the fact that practical reasoning is the cause of what it understands 'means more than that practical knowledge is observed to be a necessary condition of the production of various results [...] It means that without it what happens does not come under the description—execution of intentions—whose characteristic we have been investigating'. Obviously, then, practical knowledge is *also* a necessary condition of the production of intentional movements.

[230] Of course, there may be exceptions: it may be that someone raises her arm, and that a clever device implanted by an evil neuro-scientist would have made her arm go up anyway if she didn't have the intention. Still, reasons-explanation would be a *façon de parler* if it were generally true that reasons or intentions made no difference to the occurrence of actions.

[231] It seems that CTA would qualify as a mechanist account of action, in Anscombe's sense. For on CTA, action is produced by events that, presumably, take place in the agent's brain. Assuming that these events themselves have causal explanations in terms of other physical events (perhaps on an even lower level of description), that amounts to saying that reasons-explanation is 'comparable to shadows'. Saying that an agent is doing φ because she intends to ψ will be just an instrumentally useful way of explaining her movements. Or as I put it in §4.3.1: the agent's reasons *qua* reasons are causally impotent.

scription still supervenes (in Anscombe's sense) on their movement and interaction. However, Anscombe argues:

> [. . .] the position is not symmetrical. If you are a determinist at any level, it appears to me that you *must* be a 'mechanist' in relation to 'higher level' descriptions: you must regard them as supervenient. Thus if you are a determinist about particles and forces you must regard as supervenient the descriptions of the actions and reactions of chemical substances, and of the actions of humans and other animals. [. . .] If, however, you are indeterminist at any level, you may or may not be a mechanist in relation to higher-level descriptions. Thus determinism settles the question of mechanism, indeterminism leaves it open. [Anscombe 1983, pp. 104-105]

Anscombe's suggestion is that intentional action is incompatible with universal determinism because the latter settles the question of mechanism. If universal determinism were true, everything would have sufficient causes at the physical level, and so there would be no productive role left to play by the agent's reasoning. The issue is one of a conflict between two forms of explanation: mechanistic explanation and reasons-explanation. Indeterminism is thus necessary (but not sufficient) for reasons-explanations to be more than a *façon de parler*.[232]

I think this argument for indeterminism is convincing, given the two assumptions that reasons-explanation cannot be reduced to ordinary causal explanation, and that reasons-explanation must be more than a *façon de parler*. However, many philosophers will be inclined to reject the conclusion and insist that one of the two assumptions must be wrong. For the idea that an agent's reasoning must play a productive role—explaining why one of many physically undetermined events comes about—while not being reducible to ordinary causation might seem to imply a form of dualism. The fear is that an agent or his reasons would somehow have to exercise their causal influence from outside the natural world. A defense of an Anscombean incompatibilism will have to show this fear to be unfounded. The challenge is to show how an agent's practical reasoning can truly be one form of *causality*, among other such forms occurring in the natural world. That is what I will try to do in the next section (§5.2). I argue that an account of causality in terms of the exercise of *powers* allows us to meet the challenge. I will not attempt a full-fledged defense of powers-based accounts against other theories of causation. Instead, I will try to show how thinking in terms of the powers or capacities of objects can make room for powers that exist only on higher levels of description—of which the power of

[232] It is important to note that at this point, all we have is an argument that intentional action requires indeterminism on lower levels of description. That leaves open whether happenings as they are described on the higher-level, e.g. of action-explanation, are determined on their own level of description. So the current argument does not yet provide a full-fledged proof of incompatibilism. I explain this in more detail in §5.2.5. Also see fn. 272. I take up the task of arguing for higher-level indeterminism starting in §5.3.

practical reasoning, or the power to act intentionally, is one.

5.2 Causality, powers, and laws of nature

Until this point, I have spoken loosely of 'ordinary causal explanation' as contrasting with reasons-explanation, without inquiring what such ordinary causality exactly is. Many philosophers of action take it that ordinary causality is somehow less mysterious than reasons-explanation, and that we can thus give an analysis of the latter by reducing it to the former. Following Davidson, they often claim that ordinary causation is *event causation*: for example, the event of ball *A* hitting ball *B* caused the event of *B*'s rolling into the pocket. Traditionally, it is thought that an event is the cause of another when the two events instantiate a law of nature. A law of nature, on such a view, is understood as an *exceptionless generalization*, or what is also called a 'covering' or 'universal' law. Such a law runs, e.g., 'Always when there is an event of type *A*, an event of type *B* follows'.[233]

The worry I identified above—that if practical reasoning is an irreducible form of causality, it would have to break the laws of nature—seems to stem from this conception of event causation and the accompanying conception of the laws of nature. For if causality is exercised on the level of description of action-explanation, then it seems that surely this would have to break the regularities that exist at the lower level. Moreover, if causation is a relation between events, and if reasons are not reducible to physical events, then how can an agent's practical reasoning be causal?

However, in recent years a different approach to causality has become popular. That approach is to understand causation in terms of an object's exercise of a power, capacity, or disposition.[234] Powers-based accounts are also beginning to have impact on the free will debate, with both proponents of libertarianism and compatibilism arguing that embracing such an account favors their camp.[235]

[233] For defenders of this basic covering law model see, e.g., Goodman [1947], Davidson [1967]. This simple covering law model faces a number of problems, perhaps most importantly, that it cannot distinguish between regularities which are intuitively just accidental, and regularities which hold as a matter of law. More sophisticated accounts of universal laws have been developed in order to escape such problems. For instance, Armstrong [1983, e.g] argues that a law is a relation between universals, roughly of the form 'whenever there is *A*-ness, it necessitates *B*-ness' (against which see the famous criticism by Lewis [1983, p. 366]). Perhaps the most popular approach is that of Lewis [1973, p. 73] (followed by e.g. Earman [1978], Loewer [1996]), who argues that only the regularities which figure in a deductive system with sufficient explanatory power (the so-called 'best system') count as laws. However, in all of these cases, the assumption is that causation cannot be a purely 'local' matter (see e.g. Lewis [1994, p. 479]). That will be the important point of contention in §5.2.1.

[234] I will use the terms power, capacity, and disposition interchangeably.

[235] On the side of the libertarian, these include Mumford and Anjum [2014], Lowe [2013], Groff [forthcoming]. For compatibilists who offer a dispositional account, see for example Vihvelin [2013], Maier [2013].

In this section, I will argue that the worry that an Anscombean account of action would commit us to a queer sort of causality that runs against the laws of nature will disappear if we do not think of causation as a relation between events which instantiate an exceptionless law, but rather as the activity of a substance. A substance, approximately, is 'a persisting object . . . possessing various properties, most importantly, causal powers and liabilities' [Lowe 2008, p. 122]. Causation is the activation or manifestation of a substance's causal power or disposition.[236] Such a view is sometimes called *substance causation*, which we have already briefly seen in §1.3.2.[237] With an appropriate understanding of powers, we can shake off the idea that the 'ordinary' kind of causation exemplified by billiard balls is somehow less mysterious or more metaphysically respectable than the kind of 'causation in virtue of content' that is essential to action-explanation.

In order to properly understand the substance-causal view, I will first have to show what is mistaken about the ordinary conception of causation as the instantiation of universal laws. I do so in (§5.2.1), where I argue that causation is properly a *local* phenomenon: causation is *activity*. In §5.2.2, I then develop an understanding of the laws of nature in terms of the powers or dispositions of substances. Together, this understanding of causality and the laws of nature makes a proper notion of substance causation intelligible (§5.2.3). The resulting substance-causal picture will allow us to differentiate between deterministic and indeterministic powers (§5.2.4), and finally, to see that higher-level causation need not break the laws of nature (§5.2.5).

5.2.1 Causation as activity

I will develop my criticism of the dominant event-causal theory of causality, and the universal law account that comes with it, on the basis of Anscombe's [1971] essay on causality. Anscombe famously argues against the idea that causation is connected to necessitation and determinism. Interestingly, contemporary event-causal libertarians often appeal to this argument of Anscombe's in order to show that indeterministic event causation is possible.[238] This is ironic, because as we will see, Anscombe does not just argue that causation is not connected to determinism. She also argues that causation has nothing to with universal laws. In fact, her argument that causation need not be deterministic is *based* on the rejection of the relevance of such laws to

[236] In what follows, I will use the terms disposition and power interchangeably.

[237] In §1.3.2, I noted that I am sympathetic to the substance-causal view taken by some libertarians in the free will debate, but that to say that all causation is substance causation does not by itself explain how an action can happen for a reason. With the understanding of practical knowledge we gained in the previous chapter, I believe we can see how a specifically rational substance-causal power differs from the powers of ordinary, inanimate substances—or so I will argue in the rest of this chapter.

[238] For example, Clarke [2003, p. 33] and van Inwagen [1986, p. 4].

causation. It will thus be instructive to look at her argument for indeterminism in more detail than contemporary libertarians do.

Anscombe starts by remarking that in ordinary causal inquiries, we do not actually seem to be interested in uncovering universal laws, or discovering that some kinds of happenings invariably follow upon others.[239] But what *is* it that we want to know when we inquire after the cause of something? Anscombe suggests the following answer:

> There is something to observe here, that lies under our noses. It is little attended to, and yet still so obvious as to seem trite. It is this: causality consists in the derivativeness of an effect from its causes. This is the core, the common feature, of causality in its various kinds. Effects derive from, arise out of, come of, their causes. [Anscombe 1971, p. 136]

When we say that B is the cause of A, we say that A 'comes from' B. This may seem unilluminating, and indeed it is not meant as an *analysis* of causation, and more as a reminder of what such an analysis must capture. And, Anscombe points out, if that is what an account of causation must capture, the account in terms of universal laws does a poor job:

> . . . analysis in terms of necessity or universality does not tell us of this derivedness of the effect; rather, it forgets about that. [Anscombe 1971, p. 136]

For to say that A comes from B simply is not to say that every time there is a B, there will be an A. Equally in the other direction: to be told that every time there is a B, an A follows will allow us to derive that there was an A from the premise that there was a B. But it does not show us B as the *source* of A—it does not tell us what (if anything) it was about B that led to A. All this is clear, Anscombe continually insists, if we start with ordinary causal concepts, say, 'travel' and 'physical parenthood' [Anscombe 1971, p. 136]. To use, e.g., the latter concept is to represent something (a parent) as that from which the effect (the child) 'comes' or 'arises'. But in using the concept, we are quite obviously not referring to any universal law connecting the parent and the child. Anscombe suggests that such ordinary causal concepts are in fact more fundamental than the notion of 'causation' itself:

> . . . the word 'cause' can be added to a language in which are already represented many causal concepts. A small selection: *scrape, push, wet, carry, ear, burn, knock over, keep off, squash, make* (e.g. noises, paper boats), *hurl.* But if we care to imagine languages in which no special causal concepts are represented, then no description of the use of a

[239] As an example, she cites the case of medicine, in which we can often say that some patient caught a disease from someone else, without being able to cite a universal law to the effect that always, when you are in a room with someone else who has an infectious disease, you will catch it. [Anscombe 1971, p. 136]

> word in such languages will be able to represent it as meaning *cause*. [Anscombe 1971, p. 137]

That is, 'cause' is a general notion *abstracted* from concepts of such familiar activities. To say, e.g., that B pushed A off the table, is *already* to represent a causal relationship between them. Anscombe is thus pointing out that causal efficacy is something *local*: it is present wherever we observe something engaging in activity of the kind she mentions, regardless of what happens to A's and B's in similar setups at different places and times. Of course, more needs to be said about the sense in which such ordinary activity-concepts are causal. I will do so below in §5.2.2, where we will see that *activity* is a form of non-accidentality that pertains to the laws of nature (on a particular understanding of such laws).

Anscombe's reminder of the local character of causal efficacy strikes me as a strong reason against analyzing causation in terms of universal regularities. But dedicated event-causalists, of course, will not be convinced.[240] This is because, as Anscombe points out, their account are traditionally inspired by Hume's skepticism about the possibility of observing causation in a particular case.[241] It is important to remind ourselves of these Humean roots of universal law accounts. For contemporary neo-Humeans start from the idea that 'all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another' [Lewis 1986, pp. ix-x].[242] The laws will then be summaries of the patterns found in this 'Humean mosaic'—descriptions of what merely happened to happen. For Hume, it was clear that pointing to such regularities would not be to give an *account* of a form of non-accidentality, but to reject the very idea of a non-accidental connection in nature. Neo-Humean event-causalists, on the other hand, try to avoid this skeptical conclusion by saying that causation just is the instantiation of a universal regularity. This position is awkward because it starts with the idea that *there is no such thing* as a non-accidental connection between a particular A and B, and then goes on to try to salvage a semblance of necessity by pointing to what happens at other times and places (or even in other worlds). We need not decide whether this strange balancing act is coherent or not. It is enough to see that the rationale for the universal law account,

[240] Compare Lewis [1994, p. 479]: 'laws hold in virtue of patters spread out over all of space and time. If laws underlie causation, that means we are wrong if we think that, for instance, that the causal roles of my brain states here and now are an entirely local matter. That's an unpleasant surprise, but I am willing to bite the bullet.'

[241] Hume and his followers have argued variously that in perception, we find only one event happening, and then another—the mysterious causal connection between the two being nowhere to be found. Hume himself concluded from this premise (quite justly, it seems) that there is no such thing as causation: there is only the expectation in the human mind that like events will follow upon like, i.e., the expectation of universal regularities. But if Anscombe is right that ordinary activity-concepts are already causal concepts, then we *can* find causal efficacy in (the observation of) individual cases: 'Nothing easier: is cutting, is drinking, is purring not 'efficacy'?' [Anscombe 1971, p. 137]

[242] I already described this position in §2.1.1

and thus for event causation, ultimately comes down to a prior skeptical rejection of the possibility of causal efficacy in the particular case. If we do not start in that place, nothing stops us from adopting and developing a credible substance-causal alternative. We will now see how Anscombe's remarks about the local character of causation, together with a proper understanding of the laws of nature, result in such an alternative.

5.2.2 The laws of nature as the laws of chess

I have argued, following Anscombe, that it is a mistake to think that causation is related to the instantiation of laws, conceived of as universal regularities between (types of) events. However, that is not to say that causation is wholly unrelated to the laws of nature: it is just that we must not conceive of these laws as covering regularities between events. But what can the idea of laws of nature have to do with *substances*? Intuitively, a substance is not the kind of thing that falls under a law of nature.[243] In this section, I argue that the laws of nature are actually more closely related to the idea of substances than one might think at first sight. Again, we will take our cue from Anscombe.

Anscombe rejects the idea that laws of nature are regularities. Rather, she suggests that the laws describe a peculiar kind of properties of substances:

> Suppose we were to call propositions giving the properties of substances 'laws of nature'. Then there will be a law of nature running 'The flashpoint of such a substance is . . . ', and this will be important in explaining why striking matches usually causes them to light. This law of nature has not the form of a generalization running 'Always, if a sample of such a substance is raised to such a temperature, it ignites'. . . [Anscombe 1971, p. 138]

The propositions describing such things as flashpoints or boiling points of a substance are not universally quantified propositions. For instance, wood has the disposition to burn when heated to n degrees. But we cannot analyze this by saying that, whenever a piece of wood is heated to n degrees, it will burn up. For it may be that, for example, someone quickly extinguishes the flames so that the piece of wood is not burned up. And it would be a mistake to hedge the universally quantified proposition by saying 'wood always burns up when raised to temperature n, unless someone extinguishes the flames'. For it is by now a well-known problem for universal law accounts

[243] It is sometimes thought that substance-causal views face the so-called 'datedness objection' (see Ginet [1990, pp. 13-14] and Broad [1952, p. 215] for an early precursor). The idea is that causation is something that happens at a time, that a cause must thus happen at a particular time, and that substances are not the kinds of things which happen at a particular time. This objection can only occur to one who ignores that causation consists in a substance's exercise of a power, which does happen at at time. For more on explaining the timing of events on a substance-causal view, see §5.2.3.

that it is impossible to come up with an exhaustive list of possible exceptions to the purported universal regularity, arguably even in principle.[244] As Anscombe [1971, p. 138] already noticed, in order to get a *true* law of that kind, we would have to keep hedging the regularity with exceptions, until it ultimately becomes vacuous ('wood burns when heated to n degrees, unless circumstances are such that it does not'). Anyway, the idea that such a law could be hedged with a so-called *ceteris paribus* clause will only appeal to one who still believes that causation is tied to universal regularities.[245] If we do no believe that, it seems we should take it at face value that propositions concerning boiling points and the like do not attempt to say what *always* happens in certain circumstances.

But what kind of proposition are they, then? The answer remains somewhat implicit in Anscombe's argument. The closest she gets to a positive characterization of the laws of nature is her statement that 'the laws of nature are . . . like the laws of chess' [Anscombe 1971, p. 141]. This analogy has perhaps not received the attention it deserves. For it contains the seeds of an anti-Humean account of laws of nature that makes a satisfactory understanding of substance causation possible.

The laws of chess are propositions such as the following: the Queen can move and take diagonally, horizontally, and vertically in any direction across any number of squares. And, simplifying a bit, a Pawn can move forward one square and take diagonally—and so on for all the kinds of pieces. These laws describe, not what each piece of a certain kind will *always* do, but what they *can* do. They describe the abilities, dispositions or *powers* of the pieces.[246] Similarly, Anscombe's thought is, for 'the flashpoint of substance X is n'. That proposition ascribes a power or *disposition* to a certain kind of substance: the disposition to ignite when brought to temperature n. So Anscombe's suggestion is that the laws of nature are (or state) the dispositions of all the natural kinds that make up the 'board' of our universe.

The notion of a power or disposition requires some elaboration. I said that 'the flashpoint of substance X is n' ascribes a disposition to ignite *when brought to temperature n*. The 'when . . . ' gives the manifestation or stimulus condition for the disposition: it tells us under what circumstances the disposition manifests.[247]

[244] See Cartwright [1983] for an argument to this effect.

[245] For an overview of the debate on *ceteris paribus* law, see Earman *et al.* [2002a]. See Earman *et al.* [2002b] for an argument against *ceteris paribus* clauses. Hüttemann [2014] wants to account for *ceteris paribus* clauses in terms of dispositions. As my argument below will make clear, I think the dispositionalist need not do so.

[246] In recent years, there has been a surge of interest in dispositional or powers-based accounts of causality. See, for example, Mumford [2003], Molnar [2003], Heil [2005], Mumford and Anjum [2011], Groff and Greco [2013]. These authors all defend a form of realism about powers. It is often not noticed that Anscombe already proposed to think about causation in terms of powers. I explain the notion of powers by means of her essay because her account has the notion of indeterminism and its relation to the possibility of a non-reductive account of intentional action in view from the beginning.

[247] As I will explain in §5.2.4, *indeterministic* powers do not have manifestation conditions.

As I said, we should not take this to mean that, e.g., a wooden plank will always burn up when its temperature is raised to the flashpoint: something may prevent the manifestation. A disposition-ascription is not a universally quantified statement connecting stimulus-events with manifestation-events.[248] It does not tell us, e.g., what each and every piece of wood will always do when it is raised to a certain temperature. Rather, on the picture I am ascribing to Anscombe, it tells us what a thing does *in general*.

This generality can be seen in the linguistic form usually taken by the ascription of dispositions:[249] we say, e.g., that 'wood *burns* at *n* degrees', or that 'particles with like charges *repel* each other'. Such statements connect certain types of objects or substances with certain types of activity through a verb in the *habitual* mode.[250] Compare, e.g., 'Alice takes care of her health, so she *runs*'. This statement is not contradicted if Alice, having fallen ill despite her best efforts, decides to skip her morning runs for a week. Similarly, statements concerning the boiling point of water or the flashpoint of a match are simply not contradicted if, in a particular case, something prevents the disposition from manifesting. Thus the stimulus conditions of a disposition should not be taken as the conditions under which a certain manifestation will *always* happen, but as those under which the disposition manifests generally or *normally*, i.e., *unless something prevents it*. As Anscombe says, the idea of 'normal conditions' here 'is quite properly a vague notion' [Anscombe 1971, p. 138]: we do not need to provide a list of just when there is no special cause preventing the manifestation, as this list is in principle open-ended.

How does this appeal to the idea that what the 'normal conditions' for the manifestation are is 'quite properly a vague notion' different from the inclusion of a *ceteris paribus* clause in a universal regularity? On the regularity account, to say that a law holds *ceteris paribus* must be to say something like: always when there is a B, an A follows, if other things are equal. The 'if other things are equal' is part of the *content* of the universally quantified proposition. Therefore it must in principle be possible to fill in the *ceteris paribus* clause: only then will the law be a determinate proposition. But on Anscombe's view, the *ceteris paribus* clause is *not* part of the content of the

[248] Many (neo)-Humeans have traditionally analyzed dispositions in terms of subjunctive conditionals connecting a stimulus-event with a manifestation-event. For example, a piece of glass is disposed to break if it would break when struck. Such accounts are faced with an endless array of counterexamples in the form of so-called *masks* and *finks* [Martin 1994, Lewis 1997, Bird 1998]: circumstances which prevent the manifestation-event from occurring, even though the stimulus obtains. The problem with such analyses is that, as so often (see §2.1.1), they try to reduce the non-accidental connection between a disposition and its manifestation to putatively less mysterious non-modal connections. The non-reductive view I describe does not suffer from the same issues, as will become clear in this section.

[249] Compare Rödl [2012, pp. 186-191].

[250] Fara [2005] claims that he also gives an analysis of dispositions in terms of habituals. However, as Yli-Vakkuri [2010] argues, Fara's account is really a variant of a conditional analysis. So unlike my account, Fara's is yet another attempt to explain away dispositionality in non-modal terms.

law. A disposition is not a disposition to ψ when 'x, y, . . . and other things are equal'. Rather, that other things must be equal characterizes the kind of non-accidentality that ties a disposition or power to its manifestation. The possibility of intervention is part of the *form* of a disposition ascription, 'X's do ψ'. For example, we cannot say that it is a law of nature that, if a human being is poisoned with a certain dose of arsenic, she will die. Rather, the arsenic has the disposition to cause death: if we consider only the victim and the poison, then it will be no accident that death ensues after a suitable amount of time. When death *doesn't* ensue, there must be some further explanation of this: we then need to know what *other* substances were present, and in virtue of which dispositions they prevented the fatal result.[251] For instance, it may turn out that an antidote was administered which prevented the arsenic from killing the victim.

That a disposition may be prevented from reaching its manifestation does not mean that it remains completely idle in such cases. The proper thing to say, it seems to me, is that when the antidote was administered, the arsenic was already doing its deadly work: it was already doing damage to the nervous system, say (fortunately, the manifestation was stopped in its tracks before it was too late).[252] The manifestation of a disposition is an ongoing process, which we can describe in the progressive:[253] a wooden plank, when heated to the right temperature in a room which contains enough oxygen, will start to burn up. While it *is burning*, someone can still come along and douse the flames with a bucket of water. But if nothing intervenes, the wood will continue burning until finally it *has burned up*. Although we cannot in principle rule out that something will intervene, it will be no accident if the plank is burned up, for it is engaged in a process of burning, which is the manifestation of its disposition to burn.

5.2.3 Powers and substance causation

As Anscombe says, the laws of nature, understood as the dispositions of kinds of substances—that glass is fragile, or that water boils at 100 degrees Celsius—play an important role in explaining why a particular glass breaks, or a particular quantity of water evaporates. That this thing here is made of glass, and that glass is fragile, explains why it broke when it was struck with a hammer. That glass is fragile shows that it was *no accident* that it broke—it broke *because* it is fragile. *This* is substance

[251] Compare Anscombe [1971, p. 138]: 'Thus the conditional "If it doesn't ignite then there must be some cause" is the better gloss on [the proposition that matches burn at n degrees], for it does not pretend to say specifically [. . .] what *always* happens.'

[252] Here the analogy between the laws of nature and the laws of chess breaks: a chess piece is never *in the middle* of manifesting its power to move—it is first on one square, and then on another.

[253] See my discussion of progressive aspect in §4.3.4.

causation. We explain what happens to a particular thing here and now by saying that it is made of glass, and *glass is fragile*. Thus, an individual substance has certain dispositions that belong to it in virtue of the kind of substance it is—in virtue of its nature or *essence*, as we may say.[254] This of course presupposes a strongly Aristotelean outlook on metaphysics. But I think that the idea is appealing. On this picture, the laws of nature describe the dispositions of a substance-type N. So together, the laws governing the N's say what it is to *be* an N.[255] That is not to deny that we can sometimes give an explanation of why substances of type N come with the power to ψ, while those of type M do not: for example, we can point out that humans do not possess the power to fly because they lack wings, or we can say that iron has a higher melting point than water because the forces that tie iron-molecules are stronger. But it seems that this will be to give an explanation of a different sort than a *causal* explanation. And it is important to note that there is no reason to suppose that there must always be an explanation of why a substance has a certain power in terms of its underlying makeup: at some point, it seems, we will simply have to accept that it is a law of nature that 'the N does ψ'. Why, for example, do negatively charged particles repel each other? It is not clear that there is or *has* to be an answer to this question in terms of the powers of the building blocks of the relevant particles. And, as I will argue in §5.2.5, it may be the same with the powers of higher-level substances.

Now the dispositions of a substance determine what it can do, or what can happen with it, in particular circumstances. Given that the glass was struck and that glass is fragile, certain future events were ruled out: the glass could not, for example, keep standing on the table, or continue to contain the water that was in it. This is, again, analogous the laws of chess. Given a particular configuration of pieces on the board, the laws of chess (the powers of the pieces) determine a range of possible future moves in the game. Some moves will be ruled out from the start, while other possibilities will be ruled out as the game progresses.

This possibility-determining character of dispositions is important. Let us look at a more complicated example. A magnet has the disposition to attract metal. So if there is an appropriate piece of metal within a certain range of the magnet, it is possible that it will, within a second, say, touch the magnet. But now suppose the piece of metal is held in place by another substance, e.g., by an event stronger magnet. Then, it seems, it is not *really possible* that the piece of metal will touch the

[254] This view should not be confused with the view called 'dispositional essentialism', which is just the thesis that there are some properties which are essentially dispositional (as opposed to being ultimately reducible to non-dispositional properties). See, e.g., Bird [2007], Ellis [2009].

[255] For comparable neo-Aristotelean outlooks on metaphysics, see e.g. Wiggins [2001], Lowe [2009], Groff and Greco [2013].

first magnet within the given timeframe. Real possibility, as I understand it, is what is possible for a certain object at a place and time.[256] Now if we widen the timeframe sufficiently, to e.g. the next five minutes, it may turn out that it *is* really possible for the piece of metal to touch the first magnet: it may be possible for someone to come along and remove or deactivate the second, stronger magnet. But it need not be: if there is nobody close enough to reach the scene within five minutes, it will *not* be really possible for the piece of metal to move from where it is sitting now to the first magnet. In that case, it will, in one sense, be *necessary* that it does not move (in the next five minutes). Yet in another sense, this will be an accident: for it is just an accident that there is nobody around to intervene. This does not change the nature of the substances involved: the metal still has the disposition to be attracted by magnets (and vice-versa). As we can see, these two senses of (non-)accidentality are not independent of each other. Whether it is really possible for the piece of metal to move within the next five minutes *depends* on whether there are any other objects *close enough* which have the disposition to do certain things like removing magnets or moving about pieces of metal—for example, human beings, or clever robots. And the 'close enough' here means: whether any of the candidate objects have a potentiality to move towards the scene within the given timeframe.

Now it may seem that there are *two* things that deserve the name of substance causation. On the one hand, there is this constraining of the range of real possibilities by powers. A power constrains what is possible even when it remains passive. On the other hand, there is the *manifesting* of a power, which is the actualization of one of the prior real possibilities. It is in manifesting a power that a substance becomes causally *active*. It is the latter that one might think causality *really* consists in—the particular pushings and pullings that Anscombe rightly insisted we focus our attention on. However, it would be a mistake to emphasize the difference between these two notions of causation. *Potentiality* and *activity* are just two poles of what we may call the causal modality. When a substance engages in an activity, this is *no accident* precisely in the sense that what is happening is something that lies in the nature of the substance to do. For example, given that the liquid in the pan is water, it is no accident if it boils after sitting in the stove for a while, for *water boils at 100 degrees Celsius*. We thus explain what happens to a particular quantity of water here and now by pointing to what water in general can do.

However, a proponent of event-causalism might insist that none of this shows that we can do away with event-causality. For saying that the stuff in the pan is water does not *really* explain its boiling. The *true* explanation is rather that the stuff

[256] For a more detailed account of the notion of real possibility, based on the branching time framework, see Müller [2012].

in the pan has been heated for a while, until it reached the required temperature. And this is to identify an *event*, namely of heating, as the cause. Indeed, it is true that (at least in this case) a full causal explanation must mention not only the disposition, but also why the substance found itself in conditions under which the disposition could manifest. But on the substance-causal picture, what explains this will itself be the manifestation of a disposition—namely, the stove's disposition to transfer heat. Powers or dispositions thus have a kind of explanatory priority. That the stuff in the pan is water explains why it boils when it reaches 100 degrees Celsius—rather than doing nothing at all, for example. This may seem trivial, but remember that on the Hume-inspired picture of laws of nature as universal regularities, there is no reason *at all* why a particular quantity of water should behave in one way rather than another. For the regularities which the laws of nature sum up, on that account, will themselves be entirely accidental.

But we should grant that the fact that the stuff in the pan is water does not offer a complete explanation of why it boiled. For it does not explain *why* it was heated to 100 degrees Celsius in the first place. So we might say that substance causation, in this case, can explain *what* happens, but not why it happened *then*. To explain the timing, we must look beyond the water's dispositions and towards other substance-causes. In fact, of course, we often only cite the triggering event when giving a causal explanation, e.g.: 'the window broke because the stone hit it'. As Ryle [1949/2009, p. 50] suggests, such an explanation is elliptical for a more complete explanation: the glass broke *when* the stone hit it, because it was fragile. However, as we will see, not *all* happenings can be given such an explanation of their timing (§5.2.4).

5.2.4 Substance causation and indeterministic powers

A substance's being active, i.e., its manifesting a disposition, does not imply that it was necessary that it would perform this activity. That is why, according to Anscombe, causation is not connected to necessitation or determinism. For we may say that it is determined that a substance will ψ if and only if it is not really possible that it will not ψ. Now it is true that a substance's powers will constrain the range of real possibilities for it, as I explained above. But nothing in the idea of substance causation implies that the range of real possibilities must always be constrained down to one. The presence of a substance with certain powers may close off some real possibilities (e.g., the presence of the styrofoam may make it impossible that the glass will break within the next five minutes), while leaving many others open (different things might still happen to the glass). Compare again the laws of chess: the powers of the pieces constrain the range of possible moves, but leave open multiple

continuations of the game.

It is important to see that the possibility of indeterminism, on this picture, does *not* lie in the fact that a disposition may in principle be prevented from manifesting, but will rather depend on whether there are substances with indeterministic powers.[257] For consider that whether or not something *will will* prevent a disposition from manifesting may itself be perfectly determined. It may be determined, for example, that the cue ball will cause the eight ball to roll into the pocket, because there is no other substance close enough to knock either ball off course. The cue ball's disposition to transfer its momentum is still something that can *in general* be prevented, but such prevention may not be really possible here and now. Whether determinism or indeterminism obtains at a given region of space and given timeframe will thus depend on which substances there are and what powers they have. More specifically, indeterminism requires that there be substances which have indeterministic powers. But what is an indeterministic power?

We can draw the distinction between deterministic and indeterministic powers as follows. A deterministic power, like water's power to dissolve salt, always requires some triggering change in order to manifest. If there is salt in a quantity of water and the salt does not dissolve, it will not suddenly dissolve unless the water or the salt undergoes a relevant change—for example, if the water's temperature is increased. By contrast, an indeterministic power, like radium's power to decay, does not have a trigger. If conditions are such that a radium atom can decay, there is nothing that can be done or added to those conditions that would force the power to be triggered.[258] So arguably, an indeterministic power lacks manifestation conditions: it can manifest *whenever*.[259] In that sense, it is true to say that undetermined events are *uncaused*, in the sense that nothing causes such an event to happen just when it did. Still, in a different sense—that of substance causation—an undetermined event *does* have a cause: for example, it is no accident that some atom decayed, because it was radium, and radium has the power to decay. What is an accident is only that it decayed at that particular time.

[257] Some libertarians who appeal to a powers-based account of causation miss this point. Mumford and Anjum [2014], for example, confuse the in principle possibility of intervening on the manifestation of a power with the absence of determinism. They thus believe that the truth of a powers-based account implies libertarianism. As I explain here, that is a mistake: there is obviously such a thing as a deterministic power. If that were not the case, it would be true, but trivial, that human action is undetermined.

[258] Of course, things can be done to force a particular quantity of radioactive material to decay: we can cause a chain reaction in it. But that is not to force any particular atom to decay at any particular time.

[259] Perhaps one would like to say that indeterministic powers do have manifestation conditions: it is just that the powers can fail to manifest even when the manifestation conditions are present, and not because something is preventing or interfering with the manifestation. However, that position invites the question what the difference in 'modal force' is between powers that do and not necessitate their manifestation when they are in the right conditions—a modal force which itself cannot be explained in terms of powers. That seems like an unfortunate move for a powers-theorist.

Note that this way of characterizing indeterministic powers is different from saying that they are so-called 'two-way powers' [Alvarez 2013, Steward 2012a, Lowe 2013]. Some philosophers (especially libertarians) use that notion to distinguish between indeterministic and deterministic powers: radium has the power to decay *or* not to decay, for example, and humans may have the power to perform some action *or* not to perform it. I think this is confused: when a radium atom does not decay during some interval of time, that is precisely *not* a manifestation of its power to decay.[260] During that time, the power remained idle and un-actualized. This is especially important for a libertarian account of freedom: I will argue that the power for acting intentionally is spontaneous precisely in the sense that it need not manifest at any given time—and that is different from saying that it must manifest, albeit in either of two ways (e.g., by moving one's arm or by keeping it still).

If there are substances with indeterministic powers, then *some* real possibilities for those substances will always be left open. We can thus conclude that determinism can only occur when and where there are no substances with indeterministic powers. It is important to note that determinism and indeterminism are thus local notions: some things may be determined at a given place and within a given timeframe.[261] But although a substance-causal account thus makes conceptual room for indeterminism, it does not provide any kind of argument to say which kinds of events at which places and times—if any!—are really undetermined. Of course, most physicists believe that the ultimate building blocks of the universe in fact have indeterministic powers. If that is right, it seems that at least *some* things—minimally on the micro-physical level, but crucially, not limited to that level[262]—will always be undetermined. But the question whether human actions are ever determined remains a substantial one to which the answer is not settled just by adapting a substance-causal account. Still, the

[260] Talk of two-way powers is often vague, in that it is not made clear whether *not* ψ'ing is really seen as a manifestation of the power. On a charitable reading, then, 'two-way power' is just a confusing label for an indeterministic power. Perhaps much that two-way theorists have to say could be reconstructed in that way. However, at times the confusing thought really does seem to be intended. For example, here is how Lowe explains the idea of a two-way power to act: 'Presented with the possibility of raising my arm on a given occasion, I can either will to raise it or alternatively *refrain* from so willing by willing *not* to raise it' [Lowe 2013, p. 177]. It seems to me that it is in general not true that whenever an agent does not φ, where φ is in her power, she will not-φ. That is, her not-φ'ing may or may not something she does intentionally, and hence may or may not be an act of her will. The question whether one can still be held responsible for refraining to act is an altogether different one (for which see, e.g., Anscombe [1991b]).

[261] Universal determinism—the claim that *everything* at all times is determined—will only be true if there are at no time and place any substances with indeterministic powers. As I explained in §1.1, a local understanding of determinism and indeterminism, as opposed to focusing only on universal determinism, is essential to the free will debate.

[262] It is still sometimes believed that indeterminism is confined to the workings of nature on the lower level. To see how mistaken this idea is, recall the example of Feynman's bomb, cited by Anscombe [1971, pp. 144-145]: 'a bomb is connected to a Geiger counter, so that it will go off if the Geiger counter registers a certain reading; whether it will or not is not determined, for it is placed near some radioactive material that it may or may not register that reading'. In this example, what happens on the micro-physical level has enormous consequences for what happens on the macroscopic level.

substance-causal picture is of real use to the free will debate in two ways. For, I argue in §5.2.5, it helps us to see how we can dissolve the worry presented in §5.1.2 for an Anscombean form of libertarianism: that, since an agent's reasons (or her reasoning) cannot be reduced to physical states or events, they would have to break the laws of nature or would somehow have to operate outside of them.

5.2.5 Indeterminism and emergence

How exactly does the substance-causal picture enable us to dismiss this worry? In short, the answer is that the worry disappears once we understand intentional action as a power of a particular kind of substance—namely, human beings.[263] If there is indeed such a power to act, it would not break the laws of nature, since the laws, on this picture, just *are* the fundamental powers and dispositions of all kinds of substances. The power to act intentionally would, in some sense, be one law among others. But of course this outline of an answer will not do. For the original problem was to understand how an agent's reasons could be truly productive or efficacious without breaking the *physical* laws of nature. That is, we want to understand how a power of intentional agency, if there is such a thing, can coexist with all the fundamental dispositions that the physical sciences teach us about.

It is here that the possibility of indeterminism becomes salient again. Remember the incompatibilist thesis we adopted from Anscombe: in order for a level of description, such as that of intentional action-explanation, to be more than a mere *façon de parler*, there must be indeterminism on the level(s) below it. Now the idea is that this does not only operate as a *negative* constraint, ruling out certain forms of description as *real* explanations, but also as a *positive* enabling condition for the existence of higher-level powers. For as long as the laws governing the lower-level substances are indeterministic, there is room for things on a higher-level (e.g., a human being) to have powers that settle which of the many real possibilities that exist on the lower level become actualized—even though, of course, the human being is in some sense ultimately composed of such lower-level substances. Indeed, for something to be a substance at all arguably just is for it to have causal powers of its own. If something is not governed by its own laws, but its behavior is explained fully by the laws of its constituent parts, it seems to be just a *façon de parler* to describe it as any*thing* over and above its parts. That is what is meant, in contemporary metaphysics, with saying that there really isn't such a thing as, say, a chair, but only 'atoms arranged chair-wise'.[264]

[263] This is not intended to suggest that *only* human beings can possess this power.

[264] The claim that there are no macroscopic objects, but only, e.g., atoms arranged chair-wise is defended by, e.g., Merricks [2001]. By contrast, Elder [2011] argues for the idea that 'familiar objects' (and not

Thus thinking in terms of substance causation, rather than in terms of event causation and exceptionless laws, enables the following sketch of a picture of human freedom and its place in nature. Nature contains a plethora of different substances, from radium atoms and complex molecules to plants and human beings. Each substance has characteristic powers of its own. And each higher-level substance is composed of substances on the lower levels. The possibility of complex, higher-level substances which are not a merely *façon de parler* implies that the laws governing the powers of the lower-level substances leave more than one real possibility open: there must remain *some* real possibilities for the higher-level substance's powers to constrain further.

An example might help to illustrate the idea. Suppose that it is not physically determined what will happen to a particular amino-acid molecule at a given time: given the physical particles around, the molecule may or may not engage in a chemical reaction with a similar molecule (for example, because whether the molecules get close enough to each other to bind depends on the undetermined actions of some electrons). But suppose that the aminos are part of a plant, which is currently manifesting one of its (higher-level) powers: the power to grow. That is, the plant is engaged in a larger process of synthesizing material for use in growing a new leaf. This power of the plant may determine it that *some* aminos will bind to each other.[265] It needn't be this exact pair, and it needn't be settled in which exact way the two molecules will hit each other. Yet, that a new molecule of this or that kind will be produced may be perfectly determined when we consider what is happening on the higher level. So if these particular two aminos end up binding each other, that will be no accident relative to the higher-level description of the process they are part of—but an accident relative to only the laws of the lower-level physical substances present.

The example is, of course, entirely fictional, and no doubt greatly underestimates the complexity of what really goes on in plants. It is merely intended as an illustration of the relation between levels of description, or of parts and wholes, on the picture we are considering. It highlights the two features of that picture that are most important for our purposes. First, the picture does *not* suppose that higher-level substances *interfere* with the workings of lower-level dispositions: no laws are broken. Second, new (in this case, biological) patterns of behavior emerge from the movements of substances whose own laws only describe quite different, statistical patterns. What we cannot explain when we consider only the mass of amino-acids and electrons at a

just living beings) cannot be reduced to their constituent parts because the objects themselves play a higher-level causal role.

[265] That is, this may be determined *if* interference with the process due to, say, a disease in the plant or the sudden explosion of a nearby bomb is ruled out.

certain place and time—that some of them will end up in a new leaf—we *can* explain by considering the (biological) laws governing the plant as a whole. But on both levels, the explanation is substance-causal: we explain what happens by reference to *what* that thing is, and the powers it therefore has.

Now at this point, I do no want to get caught up in a debate about whether plants *really* have a power to grow, or whether their leaf-growing behavior just supervenes (in Anscombe's sense) on the dispositions of the chemical and physical substances it is composed of.[266] Rather, the point is just this: that *whether* plants have powers of their own depends on whether their composite material leaves open real possibilities for the plant itself to further constrain. The same might be said about the relation between chemistry and physics: *if* the laws governing molecules are more than just *facons de parler* (which may of course still be of great instrumental use), and thus if molecules are more than, e.g, 'atoms arranged water-wise', then the laws governing their composing parts must not settle every detail of what will happen.

But doesn't the account propose a form of *emergence*? To that charge I must plead guilty. The behavior of higher-level substances is not explained by the laws governing their composite, lower-level substances, but is governed by their own laws, i.e., they have their own powers. That just is emergence.[267] However, it seems that the account does not succumb to traditional criticisms of emergentism.

The most often-cited criticism is that emergence would somehow break the laws of nature. But we have already seen that this need not be the case, as long as there is lower-level indeterminism. A related objection is that emergence would break the so-called principle of causal closure. Roughly, the principle states that all physical effects can be traced to physical causes, in a way that threatens the idea of mental causation (or indeed, any kind of higher-level causation that is not reducible to physical causation). Unfortunately, there is no single canonical statement of the principle. But witness these two formulations by prominent physicalist philosophers of mind:[268]

[266] However, for a convincing argument to the effect that the operations of living substances are not just supervenient, and thus require lower-level indeterminism, see Mulder [2016].

[267] Compare Broad's definition of emergence:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not the same kind as R; and that the characteristic properties of the whole R(A,B,C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A,B,C). [Broad 1925, p. 61]

[268] Sophie Gibb [2015, p. 628] provides a helpful list of statements of the principle of causal closure found in the literature. Most seem similar to Papineau's. Gibb [2015, pp. 630-631] also argues that these

- All physical effects have sufficient physical causes. [Papineau 1998, p. 375]
- Any cause of a physical event is itself a physical event—that is, no non-physical event can be a cause of a physical event. [Kim 2005, p. 50]

We should note that neither of these formulations is itself a law of nature, or a result of physical inquiry. Papineau's formulation is rather an *a priori* pledge of allegiance to universal determinism, while Kim's principle just seems to be the negation of emergence. As an argument against the view of substance causation developed in this section, both principles are thus merely circular. In fact, the argument against mental causation that starts from Papineau's principle just supports the idea that, *if* determinism were true, higher-level descriptions would be mere *façons de parler*.

A second criticism of emergentism could be that, even if the emergent causes do not break the lower-level laws, they are still somehow *inexplicable*. The worry seems to be that it is a mystery *how* such emergent causes could operate. However, this presupposes that causation on the lower-level of physical objects and properties is somehow *less* mysterious than causation on the higher-level. What is behind this presupposition? It may be the legacy of an event-causal picture. For suppose that causation is the instantiation of laws which are exceptionless generalizations connecting events of two types. Then it may well seem that the only things that conform to such exceptionless generalizations are 'merely physical' objects—the behavior of the ordinary objects which they compose, such as plants, chairs, and humans, is much too complicated and riddled with apparent exceptions to rules to bring under any such laws. If this is what is behind this criticism of emergence, then it seems we have identified substance causation as another condition of possibility of emergence.

Alternatively, the desire to give primacy to physical, lower-level causation may just be an expression of mechanist dogma. In that case, it need not worry us at all. For the substance-causal account precisely allows us to make sense of non-mechanistic forms of explanation. There is simply no reason why *only* small, 'merely physical' objects should possess powers. Indeed, as I have explained, to deny that things like water, plants, matches, and humans have powers of their own is arguably just to deny that they are substances at all. A philosopher who wishes to insist on mechanism may still take *that* line—but I see no reason to find it convincing.

Let us apply the picture to the subject of human freedom. On this picture, the power to act intentionally would be a higher-level power of (at least one) kind of substance: human beings. When someone acts intentionally, we can explain the actual movements that take place on a number of different levels of description: for example, physical, biological, and intentional. We can, for example, take a particular

formulations are defeated by examples of indeterministic causation.

intentional arm-movement and ask what caused it. If our inquiry is of the biological or anatomical kind, we can describe the various manifestations of dispositions of nerves and muscles. And we can zoom in further on any one of *those* happenings and ask what goes on in the nerve when it causes the muscle to contract. The answer will again be in terms of the substances making up the nerve and the muscle and their dispositions. We may grant that we can keep zooming in, until perhaps we reach fundamental links in the chain. But there is no reason to suppose that what we find when we zoom in far enough will determine it that, say, the human being whose muscles we are considering will walk down the stairs. To suppose that we *could* find such a cause would only make sense on CTA: only if there were brain states type- or token-identical to intentions could we plausibly expect to find a link in the chain that would determine an agent's action.[269] Now we have already rejected CTA in the previous chapter. So if the human does indeed walk down the stairs, we have to explain this not just in terms of the dispositions of the substances that make up the human, but in terms of the powers of the human being herself. If she were not exercising her power to act intentionally, the nerves and muscles and chemicals would not have manifested their dispositions (or not in the precise way they did on the occasion).

In this way, the substance-causal picture strikes a middle ground between CTA and the two-domains picture. Intending to do something is not being in a brain state that causes you to do it, nor is it wholly unrelated to the production of your action. Rather, acting intentionally is itself the manifestation of a higher-level causal power. Also notice that, although the laws which govern the parts of the agent, at e.g., the micro-physical level of description are probabilistic, this does not mean that what happens at the higher level is just a matter of chance:

> . . . there is nothing unacceptable about the idea that 'physical haphazard' should be the only physical correlate of human freedom of action; and perhaps also of voluntariness and intentionalness in the conduct of other animals which we do not call 'free'. The freedom, intentionalness and voluntariness are not to be analysed as the same thing as, or as produced by, the physical haphazard. Different patterns altogether are spoken of when we mention them, from those involved in describing elementary processes of physical causality. [Anscombe 1971, p. 146]

This accords with with what I argued in §2.1.2: an agent's φ'ing is not an accident precisely because it happens for a reason, i.e. is a manifestation of the power to act intentionally. What is accidental on lower levels of description—relative to the parts composing the human or animal—need not be accidental when described on the higher, intentional level of description. Thus, the luck objection will not get any

[269] Compare Anscombe [1983, p. 98].

traction.

To be absolutely clear on what this substance-causal picture of human freedom does and does not claim, it will be helpful to consider one last iteration of the objection that such causation would break the laws of nature. Wouldn't one's power to act intentionally have to break the laws governing subatomic particles?[270] For these substances are governed by *statistical* laws:

> . . . quantum laws predict statistics of events when situations are repeated; interference with these, by the *will*'s determining individual events which the laws of nature leave undetermined, would be as much violation of natural law as would have been interference which falsified a deterministic mechanical law. . . . if we have a statistical law, but undetermined individual events, and then enough of these are supposed to be pushed by will in one direction to falsify statistical laws, we have again a supposition that puts us into conflict with natural laws. [Anscombe 1971, pp. 145-146]

But the proposal is *not* that human will causes individual quantum events to be always pushed in one particular direction. To explain this, Anscombe constructs the following analogy:

> Suppose that we have a large glass box full of millions of extremely minute coloured particles, and the box is constantly shaken. Study of the box and the particles leads to statistical laws, including laws for the random generation of small unit patches of uniform colour. Now the box is remarkable for also presenting the following phenomenon: the word 'Coca-Cola' formed like a mosaic, can always be read when one looks at one of the sides. It is not always the same shape in the formation of its letters, not always the same size or in the same position, it varies in its colours; but there it always is. It is not at all clear that those statistical laws concerning the random motion of the particles and their formation of small unit patches of colour would have to be supposed violated by the operation of a cause for this phenomenon which did not derive it from the statistical laws. [Anscombe 1971, p. 146]

Of course Anscombe is not saying that such a box could actually exist. Rather, the analogy is supposed to bring out that the manifestation of a higher-level power is, in some sense, multiply realizable. A plant's growing a leaf, or someone's walking down the stairs, are things that can have indefinitely many micro-physical realizations. So, for each instance of an intentional action, the underlying micro-physical story will most likely be completely different, and hence there is no reason to think that, over time, the power to act intentionally will force the underlying micro-physical events into a pattern that does not conform to the statistical laws.[271]

[270] Pereboom [2007, p. 112] seems to think so. He argues, against the mainstream agent-causal positions we have seen in §1.3.2, that agent-causation would have to break the statistical laws. The argument below rebuts his worry.

[271] See Buchak [2013] for a similar argument, which she employs as a means to rebut the replay argument.

If we accept the substance-causal picture, indeterminism is thus arguably a requirement for *any* higher-level power in humans, animals, plants, and even inorganic substances. However, in the first two chapters, I argued that any reason to insist on incompatibilism in the free will debate would have to lie in the nature of intentional action itself. How do these two claims cohere? The answer is that there is still wiggle room for the compatibilist even on the substance-causal picture. Higher-level substance causation requires that there are some things not settled by the dispositions of lower-level substances. But might it not be determined (excluding the possibility of outside intervention) that a plant will grow a leaf at some point in time, when it is exposed to enough sunlight? The plant's dispositions, then, are responsible for ensuring that the matter of which it is composed stays together in a plant-wise arrangement, rather than, say, just falling apart. Yet *that* it would stay together in such an arrangement could still be perfectly determined. That is, there could be higher-level determinism even though there is lower-level indeterminism.[272] That is to say, it could be that a plant's power to grow is a deterministic power: given how things on the relevant level are here and now, it may be determined that the plant will grow a leaf, even though it is left open, e.g., exactly where the new leaf will form. The lower-level powers will still leave open various possibilities, but on the level of description of the plant itself these are further closed down to one. The claim that the same is true for the power of intentional action would be a higher-level form of compatibilism.

But it would be a very poor kind of incompatibilism that held it that although the micro-physical details of an agent's movements may be undetermined, it is still perfectly determined that she would, say, intentionally start walking down the stairs at t. However, for all that has been said about substance causation so far, that might be exactly how things stand with the power to act intentionally: the power to act intentionally might have a trigger. Even if that does not happen to be the case for us human, that might just be an accident: nothing has yet been said to show that it is *impossible* for an exercise of the power to act intentionally to be determined on the higher level. So substance causation by itself is not enough to secure the truth of the thesis of (Incompatibilism), as formulated in §1.1.

[272] Anscombe also acknowledges that this is a theoretical possibility, although she rightly adds that this can offer no reason to assume determinism prior to further investigation:

> . . . it must be acknowledged that the position sketched is *possible*: at the macroscopic level determinism *may* hold in some immensely complicated fashion . . . there is no need to regard the causal histories of human dealings as supervenient descriptions: we *must* do this only on the basis of a radical physical determinism. Nothing is settled by this, however, as to the possibility of holding deterministic views in relation to 'human' causality. That question is left entirely open. [Anscombe 1983, pp. 105-106]

To defeat this kind of higher-level compatibilism, we thus need to show that there is something in the very nature of the power to act intentionally that rules out such higher-level determinism. So although a substance-causal account allows us to avoid both reductionism about reasons (i.e., CTA) and the two-domains picture, a robust argument for incompatibilism must still proceed from the rational, self-conscious nature of intentional action. I provide such an argument in §§5.3-5.4.

5.3 The power of self-movement

Suppose one agrees that CTA is false, should be replaced with an Anscombean account of practical knowledge, and that causation is substance causation. Even then, why should one not rest content with a form of neutral compatibilism? Why should one think that the power to act intentionally *must* be an indeterministic power—while other higher-level powers, such as growth and photosynthesis, might be deterministic? In this section, I argue that what we have learned about the nature of intentional action in the previous chapters contains the materials to defeat this form of compatibilism. Doing so will at the same time help us to see why intentional action deserves to be called *free* action. In short, what motivates the thought that the power to act intentionally is different from, e.g., the power to grow, is the idea that acting intentionally is the specifically human form of *self-movement*. Although I will argue in favor of this idea, it will be instructive to first critically examine a recent similar attempt to argue for libertarianism by Steward [2012]. Steward wants to defend a form of agency-incompatibilism: the thesis that acting itself (and not a special kind of 'free' acting) is incompatible with determinism. It will not be a surprise that I am sympathetic to this position. However, there are some important differences between Steward's account and the one I favor.

Steward explicates the thesis of agency-incompatibilism as follows: 'if there are self-moving animals, the future is open' [Steward 2012a, p. 12]. That is, self-movement implies that universal determinism is false. But what is the link between animal self-movement and indeterminism supposed to be? As Steward notes, it is not an easy task to spell this out:

> . . . it may quite reasonably be asked why one should think that there is any relationship at all between self-movement and the open future. Surely there are all sorts of self-moving entities (cars? robots? planets?) whose operations have no particular tendency to subvert whatever confidence we might antecedently have had in determinism. Why should self-moving *animals* be different? [Steward 2012a, pp. 13-14]

A part of the answer, Steward suggests, is this:

> On one intuitive understanding of self-movement, an entity is a self-mover if and only if it can move *by* itself—that is to say, without being pushed or pulled into motion by some sort of external source, in particular by another object. [. . .] we are able to make a crude division of entities into those that never appear to move from a position of rest except when made to do so by an external impetus of some sort [. . .] and those that may do so [. . .]. [Steward 2012a, p. 15]

It may seem that if something can move from a position of rest without being pushed or pulled by an external source, it follows that its movement was undetermined. However, as Steward notes, this is problematic: must a robot, a self-driving car, or a single-celled organism be made to move by something outside of it? In a perfectly intelligible sense, the answer is no: a paramecium, Steward argues, moves by itself in a way that, say, a bicycle does not. It does not need to be picked up or literally pushed around by another entity in order to move. So the relation between a stimulus and the paramecium's movement is interestingly different from the relation between foot and bicycle, or between two billiard balls when the one makes the other roll. Yet it seems that things that move by themselves, in this sense, need not move indeterministically: why might it not be perfectly determined in what direction a paramecium will move when it encounters an obstacle, or how a self-driving car will move once someone provides it with a destination? At least an opponent of agency-incompatibilism will certainly say that that is how things stand.

Steward rightly concludes that, if there is a notion of self-movement that requires indeterminism, it must be stronger than the distinction between things that move by themselves and things that need external pushings or pullings. She suggests that this stronger conception is the following:

> . . . an entity is a self-mover if and only if it is able to *make itself (or parts of itself) move*. [. . .] There can be entities which can move *by themselves* yet cannot *make themselves move*. The paramecium, I think, would be an example. [. . .] It makes a contribution, of a kind, to its own progress through the world, in virtue of the fact that at least some important parts of the processes which cause it to respond to such things as obstacles, detected sources of light, food, warmth, etc. are internal to the cell that constitutes it. [. . .] But, I would maintain, it does not make itself move. [Steward 2012a, p. 15]

It is difficult to see how the distinction between 'moving by oneself' and 'making oneself move' helps to identify a stronger sense of self-movement. Steward adds that 'it is only of some sorts of entity that it makes sense to say that they 'have' bodies, thereby separating what is moved (a body or a body part) from what is doing the moving (an animal)' [Steward 2012a, p. 16]. In the case of the single-celled paramecium, it may seem that there is indeed no distinction between the organism and that which moves. But in the case of a self-driving car the distinction between

the car and the parts that are moved (the wheels) is very clear. Why does that not count as 'making oneself move'? Steward's answer seems to be that the movement of the car, just as that of the paramecium, is fully explained by the workings of its parts:[273] there is no 'irreducible role to be played in the explanation of [the entity's] activity by a certain kind of *integration*':

> This integration is missing from the paramecium. [...] It need not be thought of as making any contribution to what occurs in respect of the movement of its body which amounts to anything over and above that of the contribution of a few simple, chemically controlled processes occurring inside the cell which *is* its body. But the same [...] is not true with respect to those creatures I want to call true self-movers. Most animals [...] are possessors of a capacity for a kind of top-down determination [...] in such a way that their contribution *does* amount to something over and above the contribution of processes inside them... [Steward 2012a, pp. 16-17]

Steward says that for an entity to be a self-mover is for it to 'make a contribution' over and above the contribution of its constitutive parts and processes. But this, it seems, is just to say that its movement must not supervene, in Anscombe's sense: the entity's movement must be the manifestation of a higher-level power, or else it would be a mere *façon de parler* to explain the movement in terms of laws governing the entity itself. If *that* is what Steward means, the notion of self-movement collapses into that of a higher-level power. And it turns out that this is indeed all that Steward has in mind when speaking of 'top-down determination' or 'making a contribution over and above the contribution of processes inside [the animal]'. This is clear from the fact that such top-down causation is, for Steward, a ubiquitous phenomenon that occurs even in simple biological structures:

> For example, a cell is a structure that, once formed, can be a source of control over the chemical processes that go on within it in the sense that laws and principles that belong to the level of the cell overtake those that belong to the level of the molecule when it comes to understanding how those lower-level processes are integrated and harmonized to serve the purposes of the cell. [Steward 2012a, p. 245]

Similar to what I have argued above, the picture here is that the cell's parts leave open various real possibilities, some of which are then actualized by the higher-level powers of the cell itself. However, we cannot then use the notion of self-movement to argue for (higher-level) agency-incompatibilism. As we have seen, all higher-level powers require lower-level indeterminism—but that does not mean that exercises of the higher-level power will itself be undetermined. And it should be clear that the indeterminism, in the case of Steward's cell, only *has* to obtain on the lower level.

[273] And that is to say: the movement of the whole car supervenes on the workings of the parts, in Anscombe's sense (see fn. 228).

That is, we have not yet been given a *reason* to suppose that the emergent higher-level powers of the cell must themselves be indeterministic, i.e., that they have no triggering conditions. For all we have heard, it may be perfectly determined that the cell will manifest one of its higher-level powers when exposed to certain stimuli (even if the precise lower-level realization of the manifestation is left undetermined). So if top-down control is all there is to self-movement, the latter does not contain any ingredients for indeterminism *on the level of description of the functioning of the whole organism*.

So top-down determination does not elucidate the sense in which higher animals do, and paramecia (whose behavior Steward admits may be determined) don't, 'mak[e] a contribution over and above the contribution of processes inside it'. This problem is made even more urgent by Steward's admission that her notion of top-down causation may even apply to certain inanimate objects—her example is that of a whirlpool, whose operations, she submits, cannot be reduced to that of the arrangement of water molecules that constitutes it [Steward 2012a, pp. 241-243]. The lower-level arrangement leaves open several physical possibilities, some of which are realized by the whirlpool as a whole. Although I am not convinced that the workings of a whirlpool are a good example of higher-level powers, this general picture accords with what I have said about higher-level substances: even the powers of inanimate higher-level substances require lower-level indeterminism. Yet of course, the powers of many inanimate substances are perfectly deterministic. This makes it especially clear that Steward's notion of top-down determination does not really give us the 'stronger' sense of self-movement that Steward requires: it does not demarcate a class of beings (animals) that are self-movers in the exalted sense which, the agency-incompatibilist wants to say, implies freedom.

It seems to me that this problem for Steward's account arises because she defines agency-incompatibilism as the thesis that acting is incompatible with *universal* determinism. That is true, as my discussion of higher-level powers shows. But by considering only universal determinism, Steward fails to thematize the question of higher-level compatibilism sufficiently clearly. It seems as though she *wants* to say that there is a sense in which higher animals are, and paramecia are not, self-movers, and that it is *this* sense which requires indeterminism. But then it turns out that even the top-down determination in a paramecium implies the falsity of universal determinism, and the distinction between things that move themselves and things that do not collapses.

However, I think we *can* define a sense of self-movement that is sufficiently strong to establish a higher-level agency-incompatibilism. Steward's distinction between things that move *by* themselves and things that make themselves move does not help

to delineate that sense of self-movement. Let us understand the idea of something which 'moves by itself' as something that moves by or according to its own laws, in the Anscombean sense of a higher-level law that I have explained. In that sense, as I have argued, the class of things which move by themselves includes all higher-level substances. Yet there may be even stronger senses in which a substance may be governed by laws that are its own. I will argue that there are (at least) two. One is the teleological unity formed by the powers of living things. The other is the specifically rational form of self-movement exhibited in intentional action. The former, perhaps, comes close to the idea of animal self-movement that Steward seems to have in mind.[274] But I argue that the mere notion of a teleological power is in the end not enough to establish agency-incompatibilism. As we will see, that is different in the case of specifically rational self-movement.

5.3.1 Teleology and self-movement

In this section, I will argue that there is a strong sense in which living things move according to their owns laws. As we will see, that sense is intimately related to the teleological nature of the powers of living things. I explain this by a (very) short exposition of Thompson's [2008] account of life. I hope to show that the powers of living things are 'autonomous' in a way that does not apply to inanimate higher-level substances: whereas it is just an accident when an inanimate substance actualizes one of its potentialities, the actualization of a living being's powers is explained *by the lifeform itself*. However, I will also show that this form of autonomy is still far away from the freedom that is (I argue) exhibited in intentional actions—and not just as a matter of complexity.

To understand the way in which living beings operate according to their own laws, I will start with Thompson's treatment of the question what life itself is. Thompson argues against the idea that we can give a reductive definition of life in terms of some complex property possessed by all and only those substances which are alive. The reductive idea would be that we can supply a list of properties that are necessary and jointly sufficient for a substance to count as living—properties that we must, of course, be able to understand independently of the concept of life. Thompson suggests that this is impossible, because every putative sign of life will either fail to be metaphysically necessary, or will already presuppose a prior understanding

[274] This must be qualified. Steward's argument proceeds without attending to the notion of teleological explanation at all: she just does not think that the sense in which an animal or human being can act 'for a reason' has much to do with action, freedom, or indeterminism at all [Steward 2012a, p. 34]. But if my argument here is correct, it is precisely because she fails to attend to the broadly teleological nature of human and animal powers that she cannot define a notion of self-movement that does not apply to, e.g., inanimate higher-level substances.

of life.[275] DNA, for example, *might* be a feature of all living things to be found in the universe. But even if that is the case, this 'would only show how resource-poor the physical world really is. It could make no contribution to the exposition of the concept of life [. . .] except perhaps as pointing to a few gorillas and turnips might.' [Thompson 2008, p. 37] That is, there appears to be no reason why living beings *could* not in principle exist without DNA, utilizing some other chemical mechanism. On the other hand, such statements as 'living things respond to stimuli' may well be a necessary condition for life—but only if we already presuppose a certain notion 'responding' and 'stimuli' that rules out that, say, an avalanche 'is the "response" to excessive yodeling' [Thompson 2008, p. 39]. The notion of 'responding' will thus depend on an understanding of the very thing we are trying to define: life.[276]

Thompson thus concludes that we should approach the phenomenon of life in a different way. He likens his strategy to Anscombe's approach in the explanation of intentional action. As we have seen (§4.3.1), Anscombe rejects the idea that intentionality is an 'extra feature' of some events (bodily movements). Similarly, Thompson rejects the idea that life is an extra feature of some substances. And, analogously to Anscombe, Thompson argues that we should instead try give an analysis of the way in which we make representations, or form judgments, about the phenomenon in question. For Anscombe, this form of judgment is the one exhibited in the answering of her question 'Why?', and as we will see, Thompson also uncovers a sense of 'Why?'—a special form of explanation—that is essential to life. His starting point is the peculiar kind of judgments we make when describing certain species or *lifeforms*, which he calls 'natural-historical judgments' [Thompson 2008, p. 64]. Here are some examples:

- The bobcat breeds in spring.
- Horses have four legs.
- This oak tree sheds its leafs in autumn.

These judgments all describe, or give content to, what it is to be a certain lifeform. As Thompson notes, these judgment have a particular logical form that is irreducible to either Freagean universal quantification (not *all* bobcats breed in spring) or statistical generalization. For although it is a true natural-historical judgment about frogs that, say, 'tadpoles grow legs', the vast majority of tadpoles will not grow legs: they will be eaten by predators before they get the chance. These judgments thus have a certain

[275] As Thompson says: '. . . *every* candidate list-occupant must strike the sub-metaphysical Scylla of "DNA" or else sink into the tautological Charybdis of "organs".' [Thompson 2008, p. 39]

[276] Thompson argues similarly for various other putative list-occupants, such as 'living things are highly organized' or 'have low entropy'.

generality in common with the disposition-ascriptions we have seen above (§5.2.3). Just as it is true that 'iron melts at 1538 degrees', even if (absurdly[277]) no iron has ever actually melted or will ever melt, so the truth of a natural-historical judgment does not depend on whether any number of members of a lifeform ever exemplifies it in actuality. So it may seem that the logical form Thompson is after is just that of disposition-ascription. And indeed, a natural-historical judgment does ascribe a power to a lifeform or individual members of a lifeform: to say, e.g., that 'horses have four legs' is to say that horses *can grow* four legs (similarly, as we have seen, the general judgment 'iron melts' can also be phrased as 'iron *can* melt').

However, there is also an important difference between these judgments and those concerning the dispositions of inanimate substances. For to say that horses have four legs is not *just* to say that horses *can* grow four. It is also to say, in an interesting sense, that horses *need* to grow four legs. That is, a natural-historical judgment involves not only the ascription of a potentiality, but also of a form of necessity. Consider the example of the bee: *bees have wings* (of this-and-this sort, size, and weight) because *bees fly*. They need the ability to fly to do all the other things that belong to bees: e.g., move from flower to flower and gather food. And they gather food because without it, of course, they would not be able to grow (or maintain) wings and fly. So, unlike the dispositions of inanimate things, the natural-historical judgments that describe a lifeform form a closed *system*: each judgment describes something that a kind of organism *does*, in our peculiar sense, *because* it plays a role in its lifeform—i.e., because it is needed for something else the organism does, which is described by a different natural-historical judgment. Thus, such judgments give application to a specific sense of the question 'Why?' corresponding to *teleological* explanation: 'the *S* does *F* because the *S* does *G*', or 'this particular *S* is doing *F* because the *S* does *G*'.

Many philosophers will be skeptical about such teleological explanation in the natural realm. But if Thompson is right, we need this form of explanation to explain the sense of natural-historical judgments, and we need such judgments in order to represent anything as living at all. For, he argues, any judgment about an organism's 'vital' activities or features (i.e., those things it has or does *qua* living being, and not just qua material object) already points towards the totality of natural-historical judgments true of a species. And, Thompson argues, the same goes even for judgments about organisms that are 'apparently purely physical' [Thompson 2004, p. 52]:

> . . . judgements as that the organism starts here and ends here, or weighs this much, must involve a covert reference to something that goes beyond the individual, namely its lifeform. It is only in light of a conception of this form, however dim that conception

[277] For a less absurd example, consider that there are so-called super-heavy metals which exist in such small quantities that we will perhaps never see them melt. Nevertheless, there is a fact of the matter about their melting point.

> may be, that you could intelligently suppose, for example, that the tentacles are not parasites or cancerous excrescences or undetached bits of waste. [Thompson 2004, p. 52]

In judging that 'here are some tentacles', one already places the things in the 'wider context' [Thompson 2008, p. 56] of the lifeform—identifying them as playing a particular role in the system of natural-historical judgments.

Yet my purpose is not to defend Thompson's account against all manner of skeptical responses. Instead, I want to draw attention to the fact that on Thompson's picture, there is a strong sense in which the vital operations of living things are no accident. I say in a strong sense, because we have already seen that when an (inanimate) substances engages in an activity, this is no accident in the sense that the activity is the manifestation of one of its powers. Given the conditions it was in, it was no accident that the iron melted, because iron has the power to melt at 1538 degrees. But the exercise of the powers of living things are no accident in a further sense. When an organism actualizes one of the powers represented in a natural-historical judgment, it is *also* no accident that it was in the right conditions to manifest that power. Consider Thompson's example of the reproduction of a plant:

> An organism's coming to be in such circumstances as tend to its reproduction *is itself typically a vital operation*, or a phase in a life-process, and therefore, in a certain sense, 'no accident'. A dandelion seed's falling on reproductively apt soil may seem fortuitous, but its parent [. . .] makes such an accident no accident, most obviously by producing so much seed. [Thompson 2008, p. 52]

In this example, the non-accidental relation is between a seed's power to grow into a new dandelion and its parent's power to produce seed—i.e., between two distinct members of a certain lifeform. But the same is true for the exercise of different powers of a single member of a lifeform. For example, it is no accident that a plant is now engaging in photosynthesis: it has the power to do so when there is enough water and sunlight. But *that* there is enough water and sunlight is also something that is provided for by the lifeform: the plant has roots to take water from the soil, and grows in such ways as to maximize its exposure to sunlight. And, of course, the power for photosynthesis in turn plays a role in the maintenance of roots and leaves. So both *what* happens in certain circumstances, and *that* the organism is actually in those circumstances, is explained by the same thing: its lifeform. If all goes well, the lifeform ensures that such things as the release of seeds (in a plant) or the growth of legs (in a frog) happen at the *right* time.

This makes for a fundamental contrast between teleological powers and the powers of inanimate substances. For although the fact that a particular iron bar just started to melt is explained by the power of iron to melt at 1538 degrees, the iron

does *not* explain why it came to be heated to 1538 degrees. It is an accident, as far as the iron is concerned, that it melts there and then—or even at all. Unlike the laws describing the powers of a lifeform, the laws governing inanimate kinds do not explain why a particular substance of a kind exercises one of its powers *when* it does.[278] So when a living thing engages in a vital operation, its activity is *its own* in a sense that the melting of the iron bar is not: the origin of the change or movement, in the first case, lies fully within the organism itself.[279]

As we have seen, Steward looked in vain for a sense in which animals move themselves, rather than being moved by an external source. This account of the teleological nature of the powers of living things arguably supplies the required notion of non-externality. An organism moves itself if its lifeform—the system of laws, represented in natural-historical judgments, that governs it—ensures that it is no accident that the movement occurs there and then.[280]

5.3.2 Is teleology enough?

That self-movement, understood as above, can reasonably be called a step upward in 'the great chain of agency, activity, autonomy, or spontaneity' [Thompson 2008, p. 46] is illustrated by Sebastian Rödl. Rödl departs from Kant's distinction between laws of heteronomy and laws of autonomy:

> Die Naturnotwendigkeit war eine Heteronomie der wirkenden Ursachen; denn jede Wirkung war nur nach dem Gesetze möglich, dass etwas anderes die wirkende Ursache zur Kausalität bestimmte. [Kant 1999b, p. 446]

The necessity found in (inanimate) nature is 'a heteronomy of efficient causes' because every effect (manifestation of a power, in our terms) accords to the law or rule (*Gesetz*) that something other (*etwas anderes*) than the efficient cause itself (the substance whose power we are considering) determines that it exercises its causality (manifests its power). Or, as Rödl explains the point:

[278] It might be possible for one of the powers of an inanimate substance to cause the manifestation of another of its powers: for example, if a vase falls on the ground and shatters, perhaps one wants to say that it has manifested its fragility because it manifested its power to fall on the ground. But then the point repeats itself: the laws governing vases do no themselves explain why it fell on the ground just there and then.

[279] That is of course not to deny that a living being can only perform its functions thanks to the powers of the lower-level substances which constitute it. However, when an organism engages in one of its vital operations, it is the whole organism which explains what happens: after all, it is the organism which further narrows down the range of real possibilities left open by the lower-level substances.

[280] Boyle and Lavin [2010, p. 182] argue for the same definition of self-movement. Of course not all the powers of a living being will be powers of self-movement. An oak tree has the power to burn, for instance—but if it exercises this power, that is not something provided for by the lifeform. 'Oak trees burn' is not a natural-historical judgment, but an ordinary disposition-ascription. This can be seen from the fact that 'oak trees burn' is not explanatorily related to any other natural-historical judgments about oak trees.

> A law of heteronomy is one according to which one thing is determined to act by another thing; that is, a law of heteronomy bears the following form: "An *N* does *A*, if an *M* does *B* to it". If, on a given occasion, an *N* is acting according to this law, then something other than it, namely a certain *M*, has solicited its act, which soliciting act of the *M* did not itself accord with the law of the *N*... [Rödl 2007, p. 118]

Autonomy, then, is for a thing to 'determine itself to causality', to paraphrase Kant. If something acts according to a 'law of autonomy', then its act...

> ...does not depend on anything not explained by its [...] own nature. That the *N* is doing *A* then is *completely* explained by itself, viz. by what it is. [...] the *N* itself, as opposed to something other than it, subjects it to the causality of the cause that acts on it. [Rödl 2007, p. 118]

While I find this distinction between autonomy and heteronomy fruitful, and agree with Rödl that laws of the living are laws of autonomy in this sense [Rödl 2007, pp. 118-119],[281] a small note on the form of heteronomous laws is necessary. Rödl says, following Kant, that such laws bear the form 'an *N* does *A*, if an *M* does *B* to it'. This seems to imply that (in our terminology) laws of heteronomy have *triggers*—e.g., 'iron melts when something heats it to 1538 degrees'. But that is to say that a law of heteronomy is necessarily deterministic (although it may remain an open question whether laws of autonomy are necessarily indeterministic). And that seems clearly false. A radium atom's power to decay, for instance, is not autonomous. The atom does not 'determine itself to causality'—*nothing* does that. The laws governing radium precisely do not explain when an atom will exercise its power to decay: the timing remains an accident. We should understand the form of laws of heteronomy so that indeterministic powers are not excluded: a law governing *N* is heteronomous if the timing of the manifestation of the power it describes is not explained by a law governing *N*.

Must a power for self-movement, governed by a law of autonomy, be an indeterministic power? Arguably not—or so an opponent of agency-incompatibilism will certainly claim. For among the organisms that are self-movers in the strong sense outlined above are not just animals, but also plants and even single-celled organisms. And, as I said above, it is not clear that, e.g., a plant's power for photosynthesis must itself be indeterministic (although, to repeat the point, lower-level indeterminism is still required). The sense in which a lifeform makes the manifestation of a power no accident, one might argue, is just that the exercise of one power leads to the exercise of another. And the exercise of the second power may be determined by the result of the first. So the sense in which the presence of sunlight and water is no accident does

[281] I also agree with him that rational action exhibits an ever stronger sense of autonomy [Rödl 2007, p. 120], as I explain below (§5.4).

not seem to imply that is not determined that the plant will engage in photosynthesis when there is enough sunlight and water (and nothing intervenes). I thus believe that to find an argument for agency-incompatibilism, we must look even higher up what Thompson calls the chain of autonomy or spontaneity. We must circumscribe an even stronger form of self-movement.

Some may wish to say that the powers of *animals* (as opposed to, e.g., plants) provide a sufficiently strong sense of self-movement. Indeed, the term 'self-movement' is often reserved for self-movement with respect to place, or locomotion. Such an argument may be possible. In fact, I find it highly doubtful that the locomotive powers of bobcats and great white sharks could be deterministic. For one thing, there just do not seem to be any true laws stating the triggering conditions for those powers.[282] But a compatibilist about animal agency will disagree. Perhaps we cannot formulate these triggering conditions precisely enough, but maybe they still exist: e.g., 'deer run away in the direction X when they see a sufficiently dangerous predator sufficiently nearby'. The precise thresholds are of course very difficult to determine, but the compatibilist will say they are there anyway. Thus a stronger argument is required. If animal self-movement indeed involves a higher form of autonomy, we must explain how it differs formally from the laws governing 'lower' forms of life.[283] That may or may not be possible (although I am sympathetic to the idea). But finding out would require a positive account of the specifically animal powers of self-movement and perception, and I confess I do not know how to do that. For that reason, I will restrict myself to arguing that specifically human, rational action constitutes a higher form of self-movement that is indeterministic. Even if we want to follow Steward and others in claiming that animal self-movement exhibits a kind of freedom (compatible with determinism or not), rational action is free in an even stronger sense.

5.4 The power to act intentionally

In the previous section, we have seen that the sense in which living things move in accordance with laws that are their own arguably does not yet provide us with a conception of self-movement that requires indeterminism at all relevant levels. For all that has been said, the compatibilist will insist, the powers of lifeforms may still be deterministic: given that this plant here is an oak tree and that oak trees produce leafs in spring, it may be determined that it will grow a leaf before soon. Regardless

[282] Of course, this is less clear when we look at less complex animals.

[283] In Thompson's [2008, p. 47] terms, that would be to say that animals occupy a 'conceptual gear' higher than plants, but lower than humans or other rational agents.

of whether this is right, I now want to argue that the same is definitely *not* true for the specifically human power of self-movement—the power to act intentionally. Like all higher-level powers, that power requires lower-level indeterminism. But, I argue, exercises of that power must also be undetermined on the higher level itself: we cannot say that, given that this is a human being, it is determined that it will φ. I will argue for this claim by returning to the topic of the previous chapter: practical knowledge. After all, the human power for self-movement is the power to act intentionally—and that, I argued, is having practical knowledge.[284]

Let us begin with a small recap. I have argued that the non-accidental connection between an agent's reasons for acting and her intentional action is of a special, *sui generis* form. When an agent is doing φ because she is doing ψ, the 'because' is that of practical reasoning: a kind of causality that is essentially self-conscious. It is essentially self-conscious, because there is no distinction between an agent's knowledge of grounds—her thought 'I am doing φ because I am doing ψ'—and the thought which causes her movement—'I should do φ because I am doing ψ'. An agent's making up her mind about what to do *is* her acquiring practical knowledge of what she is doing. Therefore, I argued, an agent's practical knowledge and its object—the intentional action she performs—are not distinct realities. Practical knowledge is the form or unity of the movements an intentionally acting agent performs. And since a thing is not distinct from its form or unity, that is to say that an intentional action is 'a thought that is a movement' [Rödl 2007, p. 19].

As we have seen (§3.4.2 and §4.3.2), the identity of a subject's knowledge with the object of that knowledge is a feature of self-knowledge generally: self-knowledge is 'knowledge by identity' [Rödl 2007, p. 124], rather than knowledge acquired by the object's causally affecting the subject. It is for this reason that I said that self-knowledge is *spontaneous*, rather than *receptive*.[285] Now I think we are in a position to see that this spontaneity of practical self-knowledge is a higher form of the autonomy exhibited in the teleological powers of living beings. In the case of teleological powers, both *what* happens and *when* it happens are explained by the organism itself, because they are explained by *what it is*—the lifeform. Therefore the origins of the manifestation of the power lie solely in the lifeform: it does not need to be acted upon by something that does not accord with its own laws. And the exercise of a

[284] It is an interesting question what the relation is between the general power for practical knowledge, and more specific powers to perform a particular action-type. Some agents have the ability to swim, for example, while others do not. How can that be if they both have the power to act intentionally? Arguably, the power to swim is a way in which one can exercise the power for practical knowledge. Since the latter power is a power to self-apply action-concepts, perhaps we can say that acquiring the power to swim is to acquire an action-concept: it is to learn to use one's power for practical knowledge in a new way, i.e. it is to acquire *know-how*.

[285] Again, see §3.4.2 for the claim that self-knowledge in general is not receptive, and see §4.3.2 for this claim as it applies to practical knowledge.

self-conscious power, I suggest, has its origins in the bearer of that power in an even fuller sense. The difference is this: whereas a (mere) teleological power is exercised *because it is the right time* for the lifeform to manifest it, a self-conscious power, such as the power for practical knowledge, is exercised *because the subject* judges *that it is the right time*[286] to manifest it. Both these formulas, we should again take care to note, state the *form* of non-accidentality typical of these two kinds of powers.

Thus, self-conscious action is a yet higher link in Thompson's 'great chain of agency, activity, autonomy, or spontaneity' [Thompson 2008, p. 46]—a stronger form of moving according to one's own laws. But why should we think that the power for this form of self-movement is indeterministic? To me, it seems that this follows from the fact that practical knowledge is not receptive. For a deterministic power is one that has a trigger: it has the form '*S* does *F* when . . . ', where the blank indicates something that *works on*, or *affects* the *S* (for example: 'salt dissolves, when it is put in water'). And as we have seen, a subject precisely does not acquire practical knowledge via some object's affecting or working on her. Thus an exercise of that power cannot be triggered. In the terms Rödl borrowed from Kant: nothing other than the subject itself determines her practical knowledge to causality.[287]

Now, in the case of the teleological powers of living things, there was still the following way of arguing that they could be deterministic. The power of, e.g., a plant, is not determined to causality by anything other than the plant itself. But perhaps the plant itself, by exercising one of its powers, may still trigger the manifestation of another of its powers. In that case, nothing from *outside* works on the plant to solicit the manifestation of the second power—but that the second power would be exercised might nevertheless have been the only real possibility. Now if I am right, this reply is not available to the compatibilist in the case of intentional action. Because judgment is essential to it, the sense in which it is no accident that an agent intentionally φ's just now is different from that of a non-rational teleological power. The power for practical knowledge, I submit, is not determined to causality by anything other than itself: not even by a previous exercise of that power.

[286] This characterization of the relevant form of non-accidentality leaves it open that the agent may wrongly judge the time to be right. That may seem to be necessary, for of course agents often intentionally do something which they should not have done. But I think that this is in fact not necessary. For one may argue that judgment essentially *is* knowledge, even if actual judgments often fall short of attaining that status. If that is right, then it is better to say that a rational power is exercised when the subject *knows* it is the right time to do so. However, nothing in the coming argument depends on this point.

[287] Contrast the case of the decay of a radium atom. As I argued (§5.3.2), the timing of that decay is fully accidental, i.e., there is no explanation of why it decayed precisely at *t*. By contrast, the power for practical knowledge is exercised at *t* because the subject judges that this is the right time to act.

5.4.1 Practical knowledge and triggers

Of course, the claim that the spontaneity of the power of practical knowledge means that it cannot have a trigger requires more explanation and defense. Even if we accept the Anscombean thesis that intentional action is characterized by practical knowledge, and even if we accept that practical knowledge is not receptive, can we not still say that it was physically determined that an agent would acquire a specific bit of practical knowledge at t? I will elucidate my thesis by arguing against this compatibilist idea.

John Hyman thinks that the power to act intentionally has, or at least could have, a trigger:

> Does the expression of desire in action invariably have a trigger? It is not *obvious* that it does, but it is plausible. I may often have no reason—i..e, no justification—for doing something at a particular moment, rather than few moments earlier or later, but it does not seem plausible that it can be a matter of pure chance. In some cases, the trigger is contact, but in more cases, it is seeing, hearing or smelling something—contact at a distance, as it were—as when a child catches sight of her mother and runs towards her. And in many cases the trigger is inside the agent's body and we do not know what it is. For example, I want to tell the children to be quiet, but say nothing, hoping they will settle by themselves. After a while, I say something. Perhaps the volume of chatter rose above a a certain threshold, perhaps my blood sugar dropped below a certain threshold. I may not know why I spoke exactly when I did. Whatever the trigger was on this occasion, it was not necessarily a thought or feeling or perception. [Hyman 2014, p. 21]

Although his terminology is slightly different than mine, Hyman also uses the term 'trigger' in such a way that a disposition with a trigger must be a deterministic disposition. He says, for example, that 'radioactive decay occurs spontaneously, without any kind of trigger' [Hyman 2014, p. 21]. And of course, if the manifestation might as well *not* have occurred at t, despite the presence of the trigger, it is unclear in what sense the trigger helps to explain why it is not 'a matter of pure chance' (i.e., an accident, in my terminology) that the manifestation occurs just at t. So even though he does not explicitly endorse the conclusion, the above commits Hyman to the thesis that intentional action is at least compatible with determinism.

What should we make of Hyman's suggestion that the timing of actions is explained by a trigger? First, we should note that there is a fundamental tension between this suggestion, and the idea that acting intentionally is having practical knowledge.[288] As Hyman admits, the trigger may be something of which the agent

[288] Indeed, it is telling that Hyman's account ignores the notion of practical knowledge. This is surprising. Hyman's stated aim is to take Anscombe's arguments seriously, and show that CTA can be saved from

has no knowledge. This arguably just is to deny the possibility of practical knowledge. For to say that an action *may* be triggered by an unknown cause is to say that acting intentionally does not yield knowledge of the *true* explanation of why the agent is doing φ. However, those with compatibilist leanings may resists this move. They may insist on the following picture. An agent's power to act manifests itself in her practical knowledge, which determines the descriptions under which the agent acts intentionally. Her practical knowledge is thus knowledge of how the things she is doing hang together rationally. Yet the timing of the manifestation of the power to act requires an explanation of a different ('causal') kind, and *this* explanation may be unknown to the agent.

Notice that the picture the compatibilist is now forced to accept is nothing other than the two-domains picture (§5.1): the compatibilist must insist that the explanation of which an agent essentially has knowledge is not an explanation of the *occurrence* of the action. That is, she must distinguish between two questions 'Why are you doing φ?'. One is the question 'Why?' in the reason-giving sense, the answer to which an agent knows practically. The other is the question 'Why?' in a 'causal' or 'triggering' sense, the answer to which is potentially unknown to the agent. And as I have already explained, this distinction between 'reasons' and 'causes' which the two-domains picture assumes is alien to Anscombe (§5.1). So I have hereby proved that accepting Anscombe's theory implies that one cannot be a compatibilist. Still, it will be useful to develop the argument that the compatibilist distinction between these two questions 'Why?' is impossible to reconcile with practical knowledge in detail. Doing so will help us to better understand the way in which the spontaneity of the human will differs from the spontaneity of teleological powers in general.

To see how the idea that the power to act has a trigger undermines the possibility of practical knowledge, let us first consider Hyman's claim that without a trigger, it would be a mere accident that an agent φ's just at the moment when she does. The timing of the action, he claims, is not explained by the agent's reasons. But that seems to be mistaken. Consider an agent who is cooking a risotto. She knows that making risotto involves first cutting onions, then (shortly) frying them, then adding the rice, and then adding some broth. Suppose we ask her: why did you add some broth to the rice? Since this is an intentional action, she will know the answer: 'because I am

Anscombean worries (e.g., about causal deviance) by adopting a powers-based ('dispositional') account of causality. According to Hyman, Anscombe is right that on a standard event-causal model, knowledge of one's reasons for acting is impossible. For if we assume a neo-Humean account of causation, an agent would be unable to discover causal efficacy in the singular case. But, he argues, 'there is no inherent difficulty in the idea that knowledge of a singular causal fact can be certain and non-inferential' [Hyman 2014, p. 10] if we adopt a dispositional account instead. Thus he pretends that the problem of how an agent can have knowledge of grounds is solved by solving the problem of how we can observe singular causation. Hyman thereby ignores the fact that practical knowledge is not only 'non-inferential and certain', but also non-observational. This fundamentally weakens his attempt to save CTA from Anscombe's criticism.

making a risotto'. Yet, we may insist, 'why are you adding the broth *just now*'? If her adding the broth is truly a part of the larger action of following the recipe for risotto, it cannot be that she has no idea of the answer. Rather, she will say: 'because I have just fried the onions and added the rice'. Having practical knowledge requires that an agent knows at which stage of her action she currently is. It is thus a mistake to think that an agent could know an explanation of *what* she is doing, but not of *when* she does it. An answer to the question 'Why are you doing φ?' is an explanation of why you are *now* doing φ.

What if we ask the agent why she started to add the broth just now *rather than*, say, a minute earlier or later, as Hyman imagines? Well, there might be an answer to that question: perhaps acting a minute earlier or later would have ruined the risotto. But Hyman is certainly right that there is often no such justification. If we would ask the agent why she did it just then, rather than a second earlier, she will probably shrug and reply by saying something like: 'for no particular reason—a second earlier would also have been fine'.[289] But Hyman is wrong to conclude that the timing of the action is therefore an accident. The reason why the agent φ'd at t is that she judged the time was right for φ'ing. That there was a window of opportunity during which the time was right does not make it an accident that she did it at t, since t is within that window. The agent's reasons for φ'ing explain the timing of her action by showing why t was *a* good time to act. Hyman's claim that this is somehow not a precise enough explanation just presupposes, without argument, that the only good explanation of an action is a deterministic one.

So a trigger is not *required* to render an action no accident in any relevant sense. And in fact, I argue, the opposite is true. If an agent's φ'ing had a trigger, it would not be intentional, and in that sense an accident. For consider again our risotto-cooking agent. According to the compatibilism we are considering, the agent has practical knowledge of what she is doing, but the fact that she would come to have this knowledge was determined. The agent's action, and thus her acquisition of practical knowledge, was triggered by something of which she is potentially unaware, e.g., by her blood sugar reaching a certain level. But this is incoherent. For it seems that this is to say that the power to act manifests itself in two things: first, the agent's representation of what she is doing, and second, the action. And then the representation will be accidental to the movement that occurs: the representation will be of an action that the agent is performing *anyway*, and not because she knows

[289] Compare Anscombe [1963, p. 25], who claims that answering the question 'Why?' by saying 'for no particular reason' is not to reject the application of that question. That is, an action can be intentional under some description X, even if the answer to the question 'Why did you X?' is 'for no particular reason'. I am suggesting that this is often an appropriate response to the question 'Why?' if one's action is described as 'doing φ at t instead of t_1' or 'doing φ instead of ψ'.

she is performing it.

To see the problem here, let us ask: of *which* action does the agent acquire practical knowledge when the blood sugar level reaches the triggering point? Is it the *whole* of the agent's subsequent risotto-making—from the very beginning, when she cut the onions, all the way to the last touch? That seems impossible. When we acquire the intention to make a risotto, we perhaps already know that we have to cut the onions, but we do not yet have knowledge of the precise movements we will perform in order to do that. So if practical knowledge were acquired at the moment at which the risotto-making action is triggered, the agent would fail to know, e.g., that the movements she is now making (moving half of the onion to the side of the chopping block) are part of the action of onion-chopping. The temporal development of the action would unfold behind the agent's back.[290] And that means she will not in fact have practical knowledge of what she is doing.

Notice that it will not help to stipulate that there is a new trigger which explains the agent's action of moving half the onion to the side of the chopping block, and her acquisition of practical knowledge of that action. For we can just repeat the point: if the agent acquires practical knowledge at the moment of triggering, that will not be knowledge of the precise way in which the action unfolds. So as the action unfolds, she will not know that the movements she is making are part of her action. A representation that one is doing φ acquired at the moment of triggering will come too early: at that point in time, there is not yet a complete action of φ'ing which the agent can represent. The action still has to unfold, and in order to have practical knowledge at all, the agent will have to know what she is doing *as* it unfolds. Practical knowledge of the movements one is making in order to do φ must be acquired *in making* those movements.

A trigger cannot explain the acquisition of practical knowledge of an action, because the trigger occurs before the action unfolds. A compatibilist interpretation of practical knowledge is thus impossible, because it makes practical knowledge accidental to the temporal development of an action. Practical knowledge is the driving force in that development: each time that our agent engages in a new phase of her action, a movement occurs *because she is making risotto*—her practical knowledge of that fact explains why she is now, say, opening the cupboard.

However, if we focus only on the productive role of practical knowledge as the driving force in an action-in-progress, it may seem that we leave open the question whether the *initial* decision of the agent to, e.g., start making risotto can have a trigger. That is: while the initiation of each phase has an explanation in the agent's practical knowledge of the overarching action she is engaged in, that action *itself*, it

[290] Compare the discussion of Lavin's argument against basic action in §4.3.4, fn. 215.

seems, cannot have such an explanation. This is apparent from the fact that practical reasoning, and thus action-explanation, has an endpoint in what Aristotle called a major or universal premise, or what Anscombe called a 'desirability characterization'. As we have seen (§4.1.1), these have the form 'such-and-such is good for humans' or 'humans need *X*'—for instance, we need *nourishment*, *shelter*, etc. Once we have reached a description of this form, the chain of questions and answers 'Why?' ends. So what, if anything, explains that an agent, at a certain point in time, starts to make risotto, rather than looking for something to drink, or engaging in home improvement for additional shelter? The compatibilist may think that, e.g., the statement 'humans need food' is a description of a human disposition: to eat *when hungry*, where the latter clause indicates a triggering condition. The explanation of why the agent adopts goal *X* at *t*, rather than doing something else, is then that these triggering conditions were satisfied.

This final compatibilist move is also mistaken. The assumption that an agent's adoption of some specific goal *X* at *t* has an explanation that falls outside an agent's practical knowledge must be false: that is apparent from the fact that an agent simply *does* have practical knowledge of the final desirability characterization of her action—we can keep posing the question why until we get to the major premise. Nevertheless, it will be instructive to consider what is wrong with the idea that a desirability characterization describes a disposition with triggering conditions.

It is, I think, correct that a major premise in practical reasoning describes a human disposition. For statements like 'humans need *X*' or '*X* suits humans' are very similar to Thompson's natural-historical judgments (§5.1.2): they generically describe *what humans do*. Whichever statements of this form are true, other than the almost tautological ones such as seeking health and shelter, they together describe what is good for beings such as us. That is why practical reasoning must end in such a generic description: practical reasoning shows what is good about a particular action by showing how it contributes to what is good in general for such as the agent is. Now, we have seen (§5.3.1) that a natural-historical judgments concerning a plant or animal is a *law* describing which powers such a lifeform has, and when it exercises them. Therefore I think it is fair to say that a true statement giving a desirability characterization describes a law that governs human beings.

These laws share some similarities with laws describing non-rational lifeforms. For instance, the fact that humans (generically) do *X* does not mean that each particular human always *will* do *X*: any individual can fail to live up to the standard provided by the law. However, in the case of plants and animals, that is just because something that does not belong to the lifeform (a forest fire, or a genetic mutation) can interfere with the proper development of its life cycle. But when such interference

is absent, the laws describing the lifeform can and do explain why the individual is doing *X*. Now, in the case of the laws governing human beings, e.g. 'humans seek shelter', more than the absence of outside intervention is required in order for an individual agent to act in accordance with it. It is, additionally, required that the agent *understands* that it is now time to seek shelter from the storm.[291] For it is not that humans (generically) seek shelter *independently* from judging that they ought to take shelter. That is, it is not that humans generally take shelter, sometimes doing so unintentionally and sometimes intentionally. Rather, acting intentionally is the *manner* in which we take shelter. So the law 'humans seek shelter' *already has the form of a major premise in practical reasoning*.

Therefore the law 'humans seek shelter (e.g. from heavy storms)' does not depict the sudden onset of a storm as a trigger for an action of running away. Rather, when the agent notices the heavy rain, she must first make a practical judgment that it is now the time to seek shelter. That judgment is the starting point for her practical reasoning—'I'm seeking shelter from the rain; there is a cave not far from here; let me go there'. True, her adopting the particular intention to seek shelter is not explained by her recognition that it is required as a means to something *else* she intends (or: instrumentally). Instead, it is explained by her recognition that, here and now, seeking shelter is what is required *tout court* (or: categorically).[292] But there is no more room for a trigger to explain why she adopts the goal than there is for a trigger to explain any of the means she takes towards it. In both cases, the explanation is the agent's knowledge that the time is right to perform the action.

Notice that again, there *may* be an answer to the contrastive question 'Why did you adopt goal *X* at *t*, *rather than* doing something else that suits humans?'—just as in the case of our risotto-cooking agent, there *may* be an answer to the question why she added the rice precisely at *t*. In the case of the risotto-cooking agent, the answer may be: a little earlier or later would have spoiled the dish. Similarly, the reason why an agent chooses to seek shelter at *t*, rather than doing *Y* which also suits humans, may be that it would have been an inappropriate time to do *Y*. However, just as in the risotto-cooking case, there *need* not be such an answer. It *may* be that adding the rice

[291] Although I have argued that a sound account of free will should not be based on considerations or intuitions about moral responsibility, it strikes me that we have here found a real connection between freedom and responsibility. The failure of a merely teleological power to manifest properly is always the result of outside interference. A tree, for example, may fail to grow sturdy roots due to some disease in its leafs. But the fact that it has this disease is, of course, not up to the tree. By contrast, the power for intentional action may fail to manifest properly *just* because of a faulty judgment of the agent's. Of course, that does not exclude that an agent may sometimes also fail to act as required due to outside interference (e.g. psychological or physiological inhibition), which is therefore excusable.

[292] For an argument that there must be room for a categorical kind of practical judgment, which is not a judgment that an action is instrumentally necessary, see Rödl [2007, pp. 38-43]. Compare McDowell [1979, §§4-5].

a little earlier or later would also have been fine, and it may be that t was an equally good time for doing Y as it was for doing X. In that case, the answer to 'Why did you do adopt X rather than Y as your goal?' may simply be: 'for no particular reason'. There is no reason why this should render her action an accident, or leave us without a sound explanation of why she did X. The thought that a sound explanation of an action must leave no alternative course of events open is, as I hope we can see by now, just an expression of the 'deterministic itch' [Anscombe 1983, p. 116] from which much contemporary philosophy suffers.

I conclude that the power to act cannot have a trigger. For acting is having practical knowledge, and an agent's practical knowledge contains the *complete* explanation of why she is acting, i.e., it explains both what the agent is doing, and the timing of the action.[293] Nothing that falls outside the agent's practical knowledge can determine her power for that knowledge to causality (to use our Kantian phrase). Rather, the power determines *itself* to causality, in the fullest sense. This is, in the end, no surprise. After all, what determines the power to causality is the agent's recognition that it is right (or that now is the time) to φ. And that recognition, as we have seen, is nothing other than the manifestation of that power: the agent's practical knowledge of doing φ.

5.4.2 Knowledge of freedom

I have argued that the power to act intentionally is an indeterministic power: as a spontaneous power, it cannot have a trigger. And so it is always a real possibility that the power will *not* manifest. Such a power, then, conforms to the thesis of (Incompatibilism) as I formulated it in §1.1. Moreover, I hope it has become clear why acting intentionally deserves the name 'free': an intentionally acting agent acts according to laws that are, in the strongest sense, her own. With this, I have almost completed my defense of the thesis that intentional action is free and undetermined action. However, I want to make clear that the argument I have presented is not just an argument to the effect that the concept of intentional action requires indeterminism. I mean that it is not just an argument for the *incompatibilism* of freedom and determinism that leaves open the question whether we, humans, are actually

[293] Of course, there are many additional explanations that we can give of why this human just made a risotto: we can, for example, explain that humans (as opposed to chimps) are capable of such actions, because they have more sophisticated motor skills, which is turn explained by certain brain structures that humans (and not chimps) have. But the point is that we do not *have* to refer to, e.g., properties of human brains in order to explain what happens—a reasons-explanation is enough to make the occurrence of the various risotto-making movements no accident. The question why humans, and not chimps, possess a power for practical knowledge is an interesting and difficult one. For one thing, we may point out that such a power arguably requires that the bearer also has certain other powers, e.g., the power to use language. But as I mentioned in §5.2.3, there is a point at which such explanations must end.

free or not. Rather, the spontaneity of practical knowledge provides an argument for full-fledged libertarianism: the thesis that we actually *are* free, and therefore, undetermined in our actions. To see why this is important, I will present a strong argument against Steward's variant of agency-incompatibilism.

We have seen that Steward argues that the concept of action requires indeterminism. She does so by arguing that the idea of indeterminism is contained in that of a self-mover. As I said, I do not think that her argument departs from a sufficiently strong notion of self-movement to warrant that inference. But even when we bracket those concerns, an important weakness remains. Consider a short version of Steward's analysis of the concept of agency:

> It seems to me quite clear—and empirical research in developmental psychology confirms it—that human beings are predisposed from a very early age to regard some of the things they meet in experience in a way extraordinarily different from the way in which they regard certain others. These special things are regarded as (i) sources of their own motion; (ii) centres of subjectivity [. . .]; (iii) targets for the application of a raft of special 'mentalistic' concepts [. . .]; (iv) possessors and controllers of things we call 'their' bodies; (v) things which are potentially suitable referents for personal pronouns [. . .] I shall call the assumptions encoded by (i)-(v) 'the agency scheme'. . . [Steward 2012b, p. 248]

Steward grants that there may be *other* concepts of agency. But, she believes, it is the agency scheme—and not the typical event-causal 'folk psychology' (i.e., belief-desire) scheme favored by many philosophers—that is *our* actual concept concept of agency. That is, the agency scheme is a 'categorisation imposed by our cognitive apparatus . . . ' which 'we have no choice but to employ' [Steward 2012b, p. 249]. Now as I said, one way to attack Steward's argument is by confronting her on her own turf: one can argue that *in fact*, the concept of agency we humans employ does not involve seeing the agent herself as the source of her own movements, or not in a sense that requires indeterminism. But even if we grant Steward that the agency scheme she describes is true to our actual psychology, it is possible to challenge the conclusion she desires: that we have libertarian free will. For, one could argue, the fact that our 'cognitive apparatus' leaves us no choice but to regard certain things in experience as sources of their own motion, does not mean that they actually *are* undetermined self-movers.[294] That is, Steward seems to give us no reason to suppose that the concept of agency objectively applies to anything at all—including to ourselves.

Notice that it is no good reply to insist that humans (or other animals) sometimes act, that acting is being a self-mover, and hence that humans are undetermined self-

[294] It seems easy to give a quasi-evolutionary or instrumental argument about why our cognitive apparatus functions in that way—for instance, thinking that the behavior of large mammals is undetermined may make one more cautious around large predators.

movers. For the current challenge is precisely to say that we know that humans *are* agents, i.e., that they sometimes *act*. Certainly, we see humans making all kinds of movements—but do any of those amount to actions, as Steward defines them? Steward may claim that it is 'clearly absurd that agency might yet turn out to be an illusion' [Steward 2012b, p. 264]. I believe that is true. But the problem is that her account does not contain any ingredients to show *why* this is absurd. It is easy to define a different concept—call it action$_D$—which is very much *like* the concept of agency except that it regards actions as determined by prior events. Who is to say that our movings around the world satisfy the concept of action, rather than that of action$_D$?

So the compatibilist may simply insist that it is an empirical question whether the actual doings of human beings are determined or not. And Steward, to an extent, agrees with that claim:

> . . . a physicist *may* come along and reveal that determinism is true. And if that were to happen, and if the physics were incontrovertible, then I should at that point have to withdraw the claim that agency is essentially an indeterministic phenomenon [. . .]. But the bare possibility of such an eventuality does not justify *now* the rejection of my thesis. [Steward 2012b, p. 248]

Two remarks are in order here. First, when Steward speaks of withdrawing the claim that agency is essentially indeterministic, it seems she must mean that actual human movements, and not the *concept* of agency as defined in the agency scheme, is deterministic. Second, she may be correct that the mere possibility of an empirical discovery that all our movements are determined does not have to undermine our conviction in the thesis that our actions are undetermined. However the problem remains: if nothing in her account rules out the possibility of this discovery, she *also* cannot justify why we should accept that thesis.

It seems that Steward thinks that, given that we actually use the agency scheme and that no physicist has incontrovertibly demonstrated the truth of universal determinism, the default position should be to trust that determinism is false. But at best, that seems like a leap of faith, and not like *knowledge* of having free will. I think this is profoundly unsatisfying for the libertarian. If we cannot offer an argument that we actually satisfy the concept of agency, it seems that all our painstaking arguments to the effect that this concept requires indeterminism would be reduced to mere wordplay. *Of course* we can cook up an incompatibilist concept of agency, if we are at liberty to define it without regard to whether anything actually satisfies it.

Moreover, I suggest, there is another reason for finding the epistemic predicament Steward leaves us in unsatisfying. She claims that the concept of agency is an empirical concept: the agency scheme is applied by our cognitive apparatus when

we encounter certain things in experience. If that is right, then it seems that when we ask whether *we* are agents, we take the same stance to *ourselves* as to other things we are confronted with in perception. Our relation to our own (purported) agency would literally be that of an observer, applying a concept to something that is *given* to us. That is, we would have to ask and *settle* the question whether some movement of ours is an intentional action or not—'here is a movement of mine, is it something I do intentionally?' And, as I have argued (§4.3.1), acting intentionally precisely means that such a question is out of order. Thus if we do not yet know that determinism is false, we cannot yet regard ourselves as agents.

I believe this shows that a satisfying libertarian account of free will cannot leave it an open question whether we are in fact agents, and thus whether we are in fact determined or not. And the argument I have presented differs from Steward's in that it does *not* leave that an open question. For, on the Anscombean account of intentional action I have argued for, the concept of agency is *not* an empirical concept. That is, it is not a concept that we predicate of an object independently given in experience.[295] Therefore it is not an empirical question whether we ourselves are agents or not.

Let me explain this. As I have argued, acting intentionally is having practical knowledge that one is φ'ing. Such knowledge, we have seen, consists in an unmediated act of self-predication of the concept 'doing φ intentionally', or what I called an action concept. And an action concept, I argued (§4.3.3), *contains* the concept of agency. To rehearse this point: an intentionally acting agent represents herself as acting *intentionally*, and since acting intentionally is having practical knowledge, that means she represents herself as having practical knowledge. In knowing that she is doing φ because she is doing ψ, the agent represents φ and ψ as things she does intentionally, and thus as things of which she has practical knowledge. So the concept of φ'ing that the agent self-applies *includes* the fact that she knows she is φ'ing. But not only that: it is also part of that concept that she is applying it *because* she knows she is applying it. That is, she not only represents herself as φ'ing, and as knowing that she is φ'ing, but she also represents the non-accidental relation between these two. Or again: her practical knowledge is a consciousness of the causality of that knowledge—and that is, consciousness of her agential activity.

Self-applying an action concept *is* self-applying the concept of agency. Therefore the concept of agency is not an empirical concept: it is not predicated of an object to which we stand in a receptive relationship, but is applied self-consciously. And

[295] That is, at least not in the first instance: the concept of agency, as I explain below, is a fundamentally a first-personal concept. But do we not sometimes judge of someone *else* that she is an agent? And if so then, then do we not apply the concept of agency to something we encounter in experience? See Rödl [2007, pp. 177-184] for a defense of a negative answer to this question. However, even if one wishes to defend that the concept of agency is an empirical concept when it is applied to others, this cannot be the case when it is *self*-applied, as I explain below.

so the question whether we are agents is not an empirical question. We do not need to wait for a scientist to come along and confirm that we are, in fact, agents. Where Steward was unable to answer the question 'how do you know that we actually satisfy the concept of agency?', we can answer it as follows: we know that we are agents *by acting*. We do not have to doubt whether a movement we make is an intentional action, because the power to act is a power for practical knowledge, i.e., knowledge that we are manifesting that very power. And since this is a knowledge that is conscious of its own causality, a subject of practical knowledge knows that this power determines itself to causality. In this way, acting intentionally *is* knowing our freedom.

There could thus be no action without representing ourselves as free. However, we should not confuse this with an idea that is common in a certain compatibilist tradition. According to this tradition, practical reasoning presupposes *viewing* ourselves as or *assuming* ourselves to be undetermined and free. But the assumption is just that: we cannot know it to be true (and it may even be false).[296] For example, according to Allison's interpretation of Kant, assuming that we are free is necessary (and sufficient) in order to deliberate practically. According to him, practical reason, as it is reason, is spontaneous. But, he argues, it is a theoretically open question whether we actually possess a capacity for *practical* reasoning, and so, whether we actually have a power for free action. For all we know theoretically, it may be that we 'act' merely on 'instinct', and so, that our actions are not free, but determined. However, for *practical* purposes, we must reject this possibility 'on the grounds that it is not a thought on which one can deliberate or act' [Allison 1997, p. 43]:

> To take oneself as a rational agent *is* to assume that one's reason has a practical application or, equivalently, that one has a will. Moreover, one cannot assume this without already presupposing the Idea of freedom, which is why one can act, or take oneself to act, only under this Idea. [Allison 1997, p. 43]

I agree that we cannot know our freedom theoretically. That is: if all we had at our disposal was the ability for theoretical, empirical knowledge, we could not demonstrate that the movements we make are free. In that case, we would be mere *observers* of our movements, trying to settle whether the concept of 'action' applies to them. But to conclude, as Allison does, that we cannot know and only *assume* ourselves to be free, one needs an additional assumption: that there is no such thing as *practical knowledge*. And, paradoxically, making that assumption positively commits one to the thesis that we are *not* free: to say that there is no practical knowledge is just to say that there is no intentional action.

[296] Apart from Allison, Von Wright [1974, pp. 133-136] can also be considered an exponent of the tradition I have in mind.

This means that the idea of freedom is only an assumption is inherently unstable. Either there is practical knowledge, and then we know we are free. Or there is not, and then it will not even be possible to *take* ourselves to act. For there is no such thing as taking oneself to act without knowing that one is acting. Taking φ to be one's action is to represent this very 'taking' as that which determines one's power to do φ to causality. A representation which does not play that causal role is simply not a representation of one's own intentional action. Thus, the idea of freedom is the idea of a kind of causality: a causality that *is* our knowledge of that causality. Being free is knowing that we are free.

* * *

Conclusion

The task I set for myself at the beginning of this thesis was to develop an account of free will that does justice to our ordinary understanding of ourselves as free agents. To do so, I said, would be to show how an action that happens for a reason is neither physically necessitated, nor an accident, but is instead an exercise of the agent's *self-determination*. Such rational self-determination, I suggested, is *spontaneous*. The main challenge in understanding how there could be spontaneous action was to see how such a *sui generis* form of causality could be part of the natural world. How can an intentional action be a spontaneous happening, if it is *qua* physical movement subject to the laws of nature? Or again: how can we determine ourself to act, if 'natural law holds sway at least over the sub-personal machinery that underlies our ability to act' [McDowell 1999, p. 102]? It is time to see how the account I developed in this thesis answers these questions.

I started, in chapter 1, by considering how the tension between accidentality and necessitation plays out the contemporary free will debate. As we have seen, the idea that a free action cannot be just an accident plays a prominent role in that debate in the guise of the luck objection against libertarianism: the worry that an undetermined action would be a matter of luck, and therefore unfree. As I argued, contemporary libertarians have no ready answer to this challenge. That is because they work with the same materials as their compatibilist opponents: the framework of event causation and the Causal Theory of Action (CTA). Even agent-causal views rely on that framework to explain how an event can be an intentional action, i.e. happen for a reason. And if we feel that an action that is probabilistically caused by a reason-state is a matter of luck, then this problem will not be solved by introducing more links of the same kind into the causal chain. On the other hand, I argued, if we accept that an agent's exercising rational control consists in causation by reason-states, then there is no reason to say that a probabilistically caused action would be a matter of luck in a sense that would render it unfree. The problem with this way out

of evading the luck objection was that it led to compatibilism: we lose the ability to say why free action should be undetermined.

As I explained, the contemporary libertarian finds herself in this difficult predicament because she fails to thematize the spontaneity of intentional action. The libertarian takes it that free action is intentional action which satisfies certain further conditions (namely, the further condition of being caused indeterministically). That is understandable, because it is difficult to see how there could be any freedom or spontaneity in an action that is caused by reason-states, which are somehow identical or correspond to states of the agent's brain. For that is something which may, after all, happen deterministically. But though it is understandable, it is the wrong response to conclude that freedom must thus consist in something we add to intentional action. As I suggested, the correct response is to insist that something that is so obviously lacking in freedom cannot be an intentional action.

However, in the contemporary debate, the possibility of arguing in that way is overlooked, precisely because it is not understood what it could be for an event to not be caused by antecedent mental states, be undetermined, and yet still not be a matter of luck. Therefore, in chapter 2, I set out to clarify what this may mean. As I explained, contemporary thinking about the tension between accidentality and freedom is hampered by what I called the uniform understanding of luck—the idea that there is a phenomenon of luck, pure and simple, which has the power to rob an agent of control over an outcome wherever it rears its head. I argued that this conception fails to deliver on its promise: the 'modal' analysis of luck it offers cannot explain in what sense accidentality undermines freedom, and in the epistemic case, knowledge.

Instead, I suggested that we should think of accidentality as the lack of an explanation of a certain form. In the case of free action, the relevant form of explanation is reasons-explanation. Being free, and that is, intentional, is a way for an event not to be an accident. Freedom is a form of non-accidentality—a form of causality. CTA claims that the relevant form of causality is just 'ordinary' causal explanation, of the kind that also governs causal transactions between billiard balls. But, I argued, this claim is optional, and not well-supported. Once we allow that there are non-accidental connections in nature (as the uniform conception of luck, with its Humean pedigree, implicitly denied), the claim that there is but one form of causality will appear optional. And, I argued, there is *prima facie* reason to suppose that CTA does not succeed in its reductive ambition, in the form of a problem of accidentality for CTA itself: the problem of causal deviance.

I thus suggested that in order to develop a sound account of freedom, we have to take seriously the idea that the kind of non-accidentality, or kind of causality,

exercised in acting for a reason—the 'because' of practical reasoning—is of a *sui generis* kind. In chapter 3, I clarified my reasons for holding that reasoning cannot be reduced to 'ordinary' event causality by focusing on the relation between self-knowledge and theoretical reasoning, i.e., reasoning about what to believe. I identified the Distinct Existences assumption (DE) as the source of the troubles that plague recent accounts of self-knowledge of belief, and argued that attempts to reconcile (DE) with an agent's *knowledge of grounds* must fail. I offered an alternative conception of self-knowledge, based on Boyle's reflectivism, which does better. On that account, to believe something *is* to self-know that one believes it, or as I said, self-knowledge is the *form* of belief. I clarified this thesis—and the reason why (DE) is false—by considering Anscombe's criticism of the idea that 'I' is a referring expression. Self-knowledge of what one believes is an unmediated act of self-predication. This self-predication is the subject's making up her mind about what to believe: it is her act of concluding that p is true, e.g., on the basis that she believes that q. Thus, I argued following Rödl [2007], reasoning is an essentially self-conscious form of causality: it is spontaneous.

Of course, there remains much work to do in developing and defending this account of the relation between reasoning and self-knowledge. For example, I have considered only the case of self-knowledge of beliefs arrived at by inference. But it is obvious that inference is not the only way of arriving at beliefs about what is the case: there are other ways, perhaps most importantly perception. More needs to be said about how we can have self-knowledge of what we perceive, and how this is possible if self-knowledge is not a 'distinct existence', i.e., not a reality distinct from what it represents. This may be an important topic for future research. However, my aim here was merely to provide a good model for thinking about what it means to say that reasoning is a spontaneous, self-conscious form of causality, in order to apply this to the practical case. This latter task I pursued in chapter 4.

In that chapter, I had two main objectives: first, to expound a non-reductive theory of action as an alternative to CTA, and second, to show why CTA is false. These two objectives, we have seen, proved to be related in interesting ways. The alternative theory of action I presented was that of Anscombe [1963]. Anscombe's theory holds that acting for a reason is essentially characterized by self-knowledge: practical knowledge of what one is doing, and of why one is doing it. As I suggested, there is thus a very clear link between Anscombe's account of action and the account of reasoning and self-knowledge I defended in chapter 3. A proper understanding of Anscombe's account should at least thematize the connection between practical knowledge and the 'unmediated agent conceptions of actions' that Anscombe mentioned in her essay on the first person. Unfortunately, this is rarely seen, and as a

result Anscombe is often misinterpreted. It is supposed, for example, that the interest of her account lies only in that she points out a familiar phenomenon—that we often or always seem to know what we do intentionally—which we then need to explain, preferably in terms amenable to CTA. That was the aim of what I called the intention-as-belief approach: to explain the phenomenon of practical knowledge in terms of the framework of mental states and 'ordinary' causal explanation.

As we have seen, that attempt fails. I argued that, if we conceive of an intention as a representation of an action which is distinct from and causally responsible for the action which it represents, then it will never represent what it must: the agent's acting *intentionally*. An intentionally acting agent represents herself as acting intentionally, and that is, for a reason. Thus she represents her reasons for acting as that on account of which she *is* acting. But that is *not* to say that she represent herself as being moved about by psychological states. For on that view, I argued, an agent would represents herself as being moved by such a cause *accidentally*—not 'through a course of practical reasoning' [Davidson 1973, p. 79]. I thus arrived at a fundamental diagnosis of the problem of deviant causal chains for CTA: there is no such thing as a 'right' causal chain, because on CTA, an agent's representation of her action never causes an action *in virtue* of the fact that it is represented. As I put it, there can be no such thing as 'causation in virtue of content' on CTA. That is why it is a false account of intentional action.

At the same time, this insight allowed us to see what a sound account of action should look like. It must not be an accident that an agent represents herself as acting, and that this representation causes her to act: as I argued, that is to say that the very act of representation must *be* the cause of why she is acting. As in the case of self-knowledge of belief, an agent's reasoning practically 'I'm doing ψ, so I should do φ' is the same act as her thinking 'I am doing φ because I am doing ψ'. It is impossible to separate the causal connection from the representation of it. This, I argued, is the correct way to understand Anscombe's account of practical knowledge. Therefore (DE) is as false in the practical domain as it was in the case of self-knowledge of belief: an intentional action is *not* a reality distinct from the agent's knowledge that she is performing it. Rather, practical self-knowledge is the *form* of intentional action.

Anscombe's account of practical knowledge thus allows us to see how an intentional action—a real-world, physical movement—may be spontaneous: the unity of some physical movements, it turns out, consists in an agent's knowledge of it. That such a movement is spontaneous follows from the definition of spontaneous knowledge as non-receptive knowledge, i.e., knowledge that springs not from the senses, but from reasoning. But does it also follow that what is spontaneous, in this sense, is also spontaneous in the sense of being self-determining, and therefore

free? And if it does, does this give us any reason to think that freedom requires physical indeterminism? It is tempting to think so, but not obviously true. And the compatibilist will resist this identification of practical self-knowledge with a form of self-determination that requires indeterminism every step of the way.

In chapter 5, I thus took on the task of arguing that in having practical knowledge, an agent is genuinely self-determining and free, in a way that implies the falsity of determinism. I first argued that the Anscombean account of action implies that universal determinism—the thesis that everything that happens is determined—must be false. This is because, as we have seen, universal determinism settles the question of mechanism (in Anscombe's [1983] sense): if the working of nature on lower levels of description were determined, then to describe an event as an intentional action would be a mere *façon de parler*. But this also raised a worry: if an agent's practical reasoning is a causal principle that cannot be reduced to what happens on such lower levels of description, then must this not be a queer kind of causation, which would arguably have to break the laws of nature?

I argued against this worry by defending a substance-causal account of causation. Causation is *activity*—the manifestation of a power of a substance. On that view, there is nothing mysterious about the possibility of higher-level causal principles. For the idea that a complex substance, such as a plant, a frog, or a human agent possesses a power is no more mysterious than the idea that, say, a billiard ball or a radium atom does so. But, I argued, this is only possible if the powers of the lower-level substances which compose, e.g., the plant are indeterministic: the parts must leave open physical possibilities for the whole to exploit with its own powers. If that is the case, then a higher-level power, such as the power to act for a reason, will not break any laws of nature. This, in effect, is my answer to the question posed by McDowell: how *we*, rational persons, can be free, even though laws of blind causality hold sway over our 'sub-personal machinery'. We can be free because the sub-personal machinery does not determine everything that happens, and so there remains room for us to determine *ourselves*. As I already noted in the introduction, this way of defending the possibility of spontaneous causality in nature has traditionally been ignored, because it is often supposed that laws of nature must be deterministic laws—and that the absence of determinism would only imply randomness. But as we have seen, this is a mistake. If universal determinism is true, then mechanism must be true, i.e., then there could be no spontaneous causal powers. On the other hand, if there is indeterminism, then mechanism *may* be false, and there is room for spontaneous causal powers.

This powers-based account, and the understanding of determinism which I argued it provides, allowed me to identify an important issue that, I think, has not yet

received attention in recent literature on free will. That is the issue that determinism may hold at different levels: it may be that, even if the powers of the substances which compose a higher-level substance are indeterministic, the powers of the higher-level substance itself are entirely deterministic. It may be determined, e.g., that a plant will grow a leaf, even if it remains open how this will exactly play out at the micro-physical level. So, *pace* libertarians like Steward [2012a], it is not enough to argue that agency implies that *universal* determinism is false. A sound libertarian account of free will which does justice to the familiar conception of agency I started with must also show that the power to act for a reason is itself an indeterministic power, i.e., a power which cannot be *triggered* by anything.

I began my attempt to provide the missing argument for higher-level indeterminism by spelling out different notions of self-movement. To move oneself, I suggested, is to move according to laws that are a substance's own, in a strong sense. In one sense, everything that a substance does is (under some description) something that accords to its own laws: everything that it can do is the manifestation of one of its powers. But, I argued, for the powers of inanimate substances, it can only be an accident that the power is manifested at the time when it is manifested. Some *other* substance must trigger the manifestation (or, in the case of e.g. a radium atom, the timing of the manifestation has no explanation and is fully accidental). This is different for the teleological powers of living organisms. With the help of Thompson [2008], I argued that a lifeform makes it no accident that its bearer exercises one of its powers at a particular time, namely, at the time when it *needs* to do so. The manifestation is then explained by nothing outside the lifeform itself. That is what the self-movement of living beings consists in.

As I said, it is not clear whether we should also differentiate on a conceptual level between the kind of self-movement exercised by, e.g., plants, and that exercised by animals. The idea that with the power of perception, a stronger form of self-movement enters the scene seems plausible. But I must leave the task of understanding animal self-movement, and how it differs from both the self-movement of plants and that of *rational* animals, for another occasion. For the purposes of this thesis, the important thing to note is that non-rational teleology still does not give us an argument to insist that a power for self-movement must be indeterministic. To prove that the power to act intentionally must be indeterministic, we therefore have to isolate a still stronger sense of self-movement.

This was the final task I set to myself. I argued that, from the self-conscious nature of the power for practical knowledge, it follows that this power cannot have a trigger, i.e., that it is an indeterministic power. For when an agent acts intentionally, she acts on account of her judgment that the time is right to act. And as we have seen

in chapter 4, this judgment—the agent's concluding that she ought to do φ—*is* her practical knowledge that she is doing φ. There is no explanation of the timing of the manifestation of the power to φ that lies outside of the agent's practical knowledge, and therefore, no trigger of the power to act. As I explained, this means that the sense in which practical knowledge is spontaneous indeed implies that in acting intentionally, an agent is genuinely self-determining and free. When we act for a reason, what happens is not determined by anything that lies outside us, nor is it an accident—*we* make it happen.

Thus, the account of intentional action in terms of practical knowledge vindicates the familiar conception of agency which I outlined at the start of this thesis. Moreover, I argued that it is not just that the concept of agency is indeed what we assume it to be on the basis of the familiar self-understanding, leaving it open whether we *actually* are agents. Rather, on the view I have developed, we can see that this familiar self-understanding embodies *knowledge*: practical knowledge of being free. For, as I explained, that the power to act intentionally is an essentially self-conscious power means that we cannot exercise it without *knowing* that we are exercising *that* power. So, I argued, whether we are free is not an open question to be decided by empirical investigation. Instead, we know that we are free by acting.

It will be useful to reflect further on this point, as it will strike many as controversial. Can a philosopher presume to make any predictions about what future science will find? Should I not admit that physicists *may* someday prove that universal determinism is true? Well, the sense in which I must admit that this is possible is the same sense in which anyone (importantly, including physicists, neuro-scientists, and free will skeptics) can be forced to admit that he *may* be wrong about some of the things he currently thinks: none of us can claim to be infallible. But although I will not labor the point against the skeptic here, from the fact that we are fallible it does not follow that we cannot know anything. So if my argument is correct, which I believe it to be, then acting freely is knowing that one is free. And thus I can say that it is *not* possible for universal determinism to be proved.

It is an interesting question why, if we can have practical knowledge of being undetermined and free, there are so many who pledge allegiance to compatibilism, or who doubt that we are free altogether. It seems to me that such a disavowal of one's freedom can only be insincere in an interesting sense: one can *say* that one disavows it, but one's actions prove otherwise. For without knowledge of the reasons of one's actions, one couldn't do *anything* intentionally, including, e.g., conducting experiments designed to disprove the existence of free will. Where does this peculiar urge to contradict one's own practical knowledge come from? I think it can only be what Anscombe [1983, p. 116] called the 'deterministic itch', or perhaps

more properly, a mechanistic itch—the idea that a sound explanation of anything, including human beings, must reveal the target of the explanation to be (the result of) a complicated piece of clockwork. That is an urge we must learn to shake off. In Ryle's words:

> The Newtonian system is no longer the sole paradigm of natural science. Man need not be degraded to a machine by being denied to be a ghost in a machine. He might, after all, be a sort of animal, namely, a higher mammal. There has yet to be ventured the hazardous leap to the hypothesis that perhaps he is a man. [Ryle 1949/2009, p. 301]

* * *

Bibliography

Allison, Henry E.

1997 "We Can Act Only Under The Idea of Freedom," *Proceedings and Addresses of the American Philosophical Association*, 71(2), pp. 39-50. (Cited on p. 225.)

Alvarez, Maria

2013 "Agency and Two-Way Powers," *Proceedings of the Aristotelian Society*, 113(1), pp. 101-121. (Cited on p. 194.)

Anscombe, G.E.M.

1963 *Intention*, Harvard University Press, Cambridge, MA. (Cited on pp. 1, 13, 20, 88, 97, 110, 122-133, 135, 151, 152, 156, 158-161, 163-166, 217, 229.)

1969 "On Promising and Its Justice, and whether it Needs be Respected in Foro Interno," *Crítica: Revista Hispanoamericana de Filosofía*, 3(7/8), pp. 61-83. (Cited on pp. 129, 131, 161.)

1971 "Causality and Determination," in Anscombe [1981]. (Cited on pp. 28, 172, 178, 183-189, 194, 199, 200.)

1974 "Practical Inference," in Anscombe [2005], pp. 109-148. (Cited on pp. 130, 131, 153, 155, 175.)

1975 "The First Person," in Anscombe [1981], pp. 21-36. (Cited on pp. 13, 75, 83, 104-108, 110-113.)

1979 "Under a Description," *Noûs*, 13(2), pp. 219-233. (Cited on p. 124.)

1981 *Metaphysics and the Philosophy of Mind. Collected Philosophical Papers Volume II*, Basil Blackwell, Oxford.

1983 "The Causation of Action," in Anscombe [2005], pp. 89-108. (Cited on pp. 178-181, 199, 201, 221, 231, 233.)

1991a *Ethics, Religion and Politics: Collected Philosophical Papers*, Wiley-Blackwell, Oxford.

1991b "The Two Kinds of Error in Action," in Anscombe [1991a], pp. 3-9. (Cited on pp. 160, 161, 194.)

1993 "Practical Truth," in Anscombe [2005], pp. 149-160. (Cited on p. 165.)

2005 *Human Life, Action and Ethics*, ed. by Mary Geach and Luke Gormally, Imprint Academic, Exeter.

Armstrong, David M.

1983 *What is a Law of Nature?* Cambridge University Press, Cambridge. (Cited on p. 182.)

Arpaly, Nomy

2009 *Merit, Meaning, and Human Bondage: An Essay on Free Will*, Princeton University Press, Princeton. (Cited on p. 153.)

Balaguer, Mark

2010 *Free Will as an Open Scientific Problem*, The MIT Press, Cambridge, MA. (Cited on p. 27.)

Bar-On, Dorit

2004 *Speaking My Mind: Expression and Self-Knowledge*, Oxford University Press. (Cited on p. 77.)

Bird, Alexander

1998 "Dispositions and Antidotes," *Philosophical Quarterly*, 48(191), pp. 227-234. (Cited on p. 188.)

Bird, Alexander

2007 *Nature's Metaphysics: Laws and Properties*, Oxford University Press, Oxford. (Cited on p. 190.)

Bishop, John Christopher
1989 *Natural Agency: An Essay on the Causal Theory of Action*, Cambridge University Press, Cambridge. (Cited on pp. 26, 27, 45, 46, 65.)

Borges, Jorge Luis
1962 "The Garden of Forking Paths," in *Labyrinths: Selected Stories and Other Writings*, New Directions Publishing, New York. (Cited on p. 21.)

Boyle, Matthew
2009 "Two Kinds of Self-Knowledge," *Philosophy and Phenomenological Research*, 78(1), pp. 133-164. (Cited on pp. 92, 112.)
2011a "'Making up Your Mind' and the Activity of Reason," *Philosophers' Imprint*, 11(17). (Cited on pp. 91, 93, 96, 97, 101, 116, 156.)
2011b "Transparent Self-Knowledge," *Aristotelian Society Supplementary Volume*, 85(1), pp. 223-241. (Cited on pp. 13, 75, 82, 94, 95, 99, 100.)

Boyle, Matthew and Douglas Lavin
2010 "Goodness and Desire," in Tenenbaum [2010], pp. 161-201. (Cited on pp. 164, 210.)

Brand, Myles and Douglas Walton
1976 (eds.), *Action theory: proceedings of the Winnipeg conference on human action*, Springer, Dordrecht.

Brandom, Robert B.
1994 *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge, MA. (Cited on p. 92.)

Bratman, Michael E.
1987 *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, MA. (Cited on p. 24.)
1991 "Review: Cognitivism About Practical Reason," *Ethics*, 102(1), pp. 117-128. (Cited on pp. 139, 140.)

Broad, C. D.
1925 *The Mind and Its Place in Nature*, Routledge and Kegan Paul, Abingdon. (Cited on p. 197.)
1952 "Determinism, Indeterminism, and Libertarianism," in *Ethics and the History of Philosophy: Selected Essays*, Humanities Press, New York, pp. 195-217. (Cited on p. 186.)

Buchak, Lara
2013 "Free Acts and Chance: Why the Rollback Argument Fails," *The Philosophical Quarterly*, 63(250), pp. 20-28. (Cited on pp. 39, 200.)

Byrne, Alex
2011 "Transparency, Belief, Intention," *Aristotelian Society Supplementary Volume*, 85(1), pp. 201-221. (Cited on p. 82.)

Byrne, Alex and Heather Logue
2009 (eds.), *Disjunctivism: Contemporary Readings*, MIT Press.

Carroll, Lewis
1895 "What the Tortoise Said to Achilles," *Mind*, 4(14), pp. 278-280. (Cited on p. 86.)

Cartwright, Nancy
1983 *How the Laws of Physics Lie*, Oxford University Press, New York. (Cited on pp. 176, 187.)

Cassam, Quassim
1997 *Self and World*, Oxford University Press, Oxford. (Cited on p. 107.)

Castañeda, Hector-Neri
1966 "'He': A Study in the Logic of Self-Consciousness," *Ratio*, 8(December), pp. 130-57. (Cited on pp. 76, 105.)

Chalmers, David J.
1995 "Explaining Consciousness: The'Hard Problem'," *Journal of Consciousness Studies*, 2(3), pp. 200-219. (Cited on p. 70.)

Chisholm, Roderick M.
1976 "The Agent as Cause," in Brand *et al.* [1976]. (Cited on p. 33.)

Clarke, Randolph

1996 "Agent Causation and Event Causation in the Production of Free Action," *Philosophical Topics*, 24(2), pp. 19-48. (Cited on p. 33.)

2003 *Libertarian Accounts of Free Will*, Oxford University Press, Oxford. (Cited on pp. 33-36, 38, 42, 68, 183.)

2005 "Agent causation and the problem of luck," *Pacific Philosophical Quarterly*, 86(3), pp. 408-421. (Cited on pp. 33, 40.)

Davidson, Donald

1963 "Actions, Reasons and Causes," in Lepore *et al.* [2006], pp. 23-36. (Cited on pp. 22-24, 60, 61, 64, 65, 173, 175.)

1967 "Causal Relations," in Davidson [2001a], pp. 149-162. (Cited on pp. 45, 66, 182.)

1969 "How is Weakness of the Will Possible?" In Lepore *et al.* [2006], pp. 72-89. (Cited on p. 25.)

1970 "Mental Events," in Lepore *et al.* [2006], pp. 105-118. (Cited on pp. 5, 23, 24, 26, 61.)

1973 "Freedom to Act," in Davidson [2001a], pp. 63-82. (Cited on pp. 26, 45, 61, 67, 69, 147, 230.)

1978 "Intending," in Lepore *et al.* [2006], pp. 122-137. (Cited on pp. 24, 25, 27, 166.)

1982 "Paradoxes of Irrationality," in Davidson [2004], pp. 169-188. (Cited on p. 25.)

1986 "A Coherence Theory of Truth and Knowledge," in Lepore [1986], pp. 307-319. (Cited on p. 88.)

2001a *Essays on Actions and Events*, Clarendon Press, Oxford. (Cited on p. 25.)

2001b "Hempel on Explaining Action," in Davidson [2001a], pp. 261-276. (Cited on p. 173.)

2004 *Problems of Rationality*, Oxford University Press, Oxford. (Cited on p. 67.)

Descartes, René

1931 *The Philosophical Works of Descartes*, ed. by Elizabeth S. Haldane and G.R.T. Ross, Cambridge University Press, Cambridge. (Cited on p. 80.)

Earman, John

1978 "The Universality of Laws," *Philosophy of Science*, 45(2), pp. 173-181. (Cited on p. 182.)

Earman, John, John T. Roberts, and Sheldon Smith

2002a (eds.), *Ceteris Paribus Laws*, vol. 57, 3, *Erkenntnis* (special issue). (Cited on p. 187.)

2002b "Ceteris Paribus Lost," *Erkenntnis*, 57(3), pp. 281-301. (Cited on p. 187.)

Ekstrom, Laura

2003 "Free will, chance, and mystery," *Philosophical studies*, 113(2), pp. 153-180. (Cited on p. 60.)

Elder, Crawford L.

2011 *Familiar Objects and Their Shadows*, Cambridge University Press, Cambridge. (Cited on p. 195.)

Ellis, Brian

2009 *The Metaphysics of Scientific Realism*, Routledge, Abingdon. (Cited on p. 190.)

Evans, Gareth

1982 *The Varieties of Reference*, 137, Oxford University Press. (Cited on pp. 74, 77, 82, 83, 85, 92, 107, 110.)

Falcon, Andrea

2015 "Aristotle on Causality," in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Spring 2015. (Cited on p. 65.)

Fara, Michael

2005 "Dispositions and Habituals," *Noûs*, 39(1), pp. 43-82. (Cited on p. 188.)

Fischer, John Martin, Robert Kane, Derk Pereboom, and Manuel Vargas

2007 *Four Views on Free Will*, Blackwell, Oxford. (Cited on p. 31.)

Fischer, John Martin and Mark Ravizza

2000 *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, Cambridge. (Cited on p. 9.)

Ford, Anton, Jennifer Hornsby, and Frederick Stoutland

2011 *Essays on Anscombe's Intention*, Harvard University Press, Cambridge, MA.

Frankfurt, Harry G.
1969 "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy*, 66(23), pp. 829-839. (Cited on pp. 20, 21.)
1971 "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, 68(1), pp. 5-20. (Cited on p. 26.)
1978 "The Problem of Action," *American Philosophical Quarterly*, 15(2), pp. 157-162. (Cited on p. 24.)

Franklin, Christopher Evan
2011a "Farewell to the luck (and Mind) argument," *Philosophical Studies*, pp. 1-32. (Cited on pp. 27-30, 46, 49, 62.)
2011b "The Problem of Enhanced Control," *Australasian Journal of Philosophy*, 89(4), pp. 687-706. (Cited on pp. 41, 42.)
2012 "The Assimilation Argument and the Rollback Argument," *Pacific Philosophical Quarterly*, 93(3), pp. 395-416. (Cited on pp. 30, 40, 54.)

Frege, Gottlob
1892 "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und Philosophische Kritik*, 100(1), pp. 25-50. (Cited on p. 103.)
1918 "Der Gedanke: eine logische Untersuchung," *Beiträge zur Philosophie des deutschen Idealismus*, I(2). (Cited on p. 103.)
1956 "The Thought: A Logical Inquiry," *Mind*, 65(259), pp. 289-311. (Cited on pp. 103, 104.)

Garrett, Brian
1998 *Personal Identity and Self-Consciousness*, Routledge, London. (Cited on pp. 107, 109.)

Geach, Peter T.
1957 "On Beliefs about Oneself," *Analysis*, 18(1), pp. 23-24. (Cited on pp. 76, 105.)

Gettier, Edmund L.
1963 "Is Justified True Belief Knowledge?" *Analysis*, 23(6), pp. 121-123. (Cited on p. 51.)

Gibb, Sophie
2015 "The Causal Closure Principle," *The Philosophical Quarterly*, forthcoming. (Cited on p. 197.)

Ginet, Carl
1983 "In Defense of Incompatibilism," *Philosophical Studies*, 44(3), pp. 391-400. (Cited on p. 176.)
1990 *On Action*, Cambridge University Press, Cambridge. (Cited on p. 186.)
1997 "Freedom, Responsibility, and Agency," *Journal of Ethics*, 1(1), pp. 85-98. (Cited on p. 30.)
2008 "In Defense of a Non-Causal Account of Reasons Explanations," *Journal of Ethics*, 12(3), pp. 229-237. (Cited on p. 152.)

Goldman, Alvin and Bob Beddor
2015 "Reliabilist Epistemology," in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Winter 2015. (Cited on p. 85.)

Goodman, Nelson
1947 "The Problem of Counterfactual Conditionals," *Journal of Philosophy*, 44(5), pp. 113-128. (Cited on p. 182.)

Grice, H.P.
1971 "Intention and Uncertainty," *Proceedings of the British Academy*, 57. (Cited on p. 128.)

Groff, Ruth
forthcoming "Sublating the Free Will Problematic," *Synthese*. (Cited on p. 182.)

Groff, Ruth and J. Greco
2013 (eds.), *Powers and Capacities in Philosophy: The New Aristotelianism*, Routledge, New York. (Cited on pp. 187, 190.)

Grünbaum, Thor
2009 "Anscombe and Practical Knowledge of What is Happening," *Grazer Philosophische Studien*, 78, pp. 41-67. (Cited on p. 164.)

Haddock, Adrian
2010 "Knowledge and Action," in Pritchard *et al.* [2010], pp. 244-328. (Cited on p. 164.)

Haji, Ishtiyaque
2000 "Indeterminism, Explanation, and Luck," *Journal of Ethics*, 4(3), pp. 211-235. (Cited on pp. 31, 36.)
2005 "Libertarianism, luck, and action explanation," *Journal of Philosophical Research*, 30, p. 321. (Cited on pp. 19, 30, 34.)

Hampshire, Stuart
1975 *Freedom of the Individual*, Princeton University Press, Princeton. (Cited on p. 175.)

Harman, Gilbert
1997 "Practical Reasoning," in Mele [1997], pp. 431-63. (Cited on pp. 136, 143.)

Heath, Joseph
2008 *Following the Rules: Practical Reasoning and Deontic Constraint*, Oxford University Press, New York. (Cited on p. 26.)

Heil, John
2005 "Dispositions," *Synthese*, 144(3), pp. 343-356. (Cited on p. 187.)

Hobart, R.E.
1934 "Free Will as Involving Determination and Inconceivable without it," *Mind*, 43(169), pp. 1-27. (Cited on p. 19.)

Hobbes, Thomas
1651/2012 *Leviathan*, Dover Publications, Mineola, NY. (Cited on p. 25.)

Horst, David
2015 "Actions and Accidents," *Canadian Journal of Philosophy*, 45(3), pp. 300-325. (Cited on pp. 55, 65, 68, 149, 154.)

Hume, David
2000 (1748) *An Enquiry Concerning Human Understanding*, Edited by Tom L. Beauchamp, Oxford University Press, Oxford. (Cited on p. 19.)

Hüttemann, Andreas
2014 "Ceteris Paribus Laws in Physics," *Erkenntnis*, 79(10), pp. 1715-1728. (Cited on p. 187.)

Hyman, John
2014 "Desires, Dispositions and Deviant Causal Chains," *Philosophy*, 89(01), pp. 83-112. (Cited on pp. 215, 216.)

Kalis, Annemarie
2011 *Failures of Agency*, Lexington, Plymouth. (Cited on p. 25.)

Kane, Robert
1985 *Free Will and Values*, State University of New York Press, New York. (Cited on p. 31.)
1996 *The Significance of Free Will*, Oxford University Press, USA. (Cited on pp. 27, 31.)
1999 "On Free Will, Responsibility and Indeterminism: Responses to Clarke, Haji, and Mele," *Philosophical Explorations*, 2(2), pp. 105-121. (Cited on p. 31.)
2002 (ed.), *The Oxford Handbook of Free Will*, 2002 edition, Oxford University Press, Oxford.
2005 (ed.), *The Oxford Handbook of Free Will*, Oxford University Press, Oxford.
2007a "Libertarianism," in Fischer *et al.* [2007], ch. 1, pp. 5-43. (Cited on p. 32.)
2007b "Response to Fischer, Pereboom, and Vargas," in Fischer *et al.* [2007], ch. 5, pp. 166-183. (Cited on p. 32.)

Kant, Immanuel
1999a *Gesammelte Schriften*, Akademie-Ausgabe, Electronic Edition published by InteLex Corp., Charlottesville (VA)., Königlich-Preussischen Akademie der Wissenschaften.
1999b "Grundlegung zur Metaphysik der Sitten," in Kant [1999a], vol. 4. (Cited on pp. 61, 210.)
1999c "Kritik der reinen Vernunft," in Kant [1999a], vol. 3. (Cited on pp. 6, 102.)

Kenny, Anthony
1976 *Will, Freedom, and Power*, Blackwell, Oxford. (Cited on pp. 175, 176.)

Kim, Jaegwon
2005 *Physicalism, or Something Near Enough*, Princeton University Press, Princeton. (Cited on p. 198.)

2007 "Contemporary Debates in Philosophy of Mind," ed. by B.P. McLaughlin and J.D. Cohen, pp. 227-242. (Cited on p. 23.)

Kitcher, Patricia

2011a *Kant's Thinker*, Oxford University Press, New York. (Cited on p. 101.)
2011b "The Unity of Kant's Active Thinker," in J. Smith *et al.* [2011], pp. 55-73. (Cited on pp. 70, 98, 102.)

Kriegel, Uriah

2003 "Consciousness as Intransitive Self-Consciousness: Two Views and an Argument," *Canadian Journal of Philosophy*, 33(1), pp. 103-132. (Cited on p. 115.)

Lavin, Douglas

2013 "Must There be Basic Action?" *Noûs*, 47(2), pp. 273-301. (Cited on pp. 122, 151, 152, 168.)
2015 "Action as a Form of Temporal Unity: On Anscombe's Intention," *Canadian Journal of Philosophy*, 45(5-6), pp. 609-629. (Cited on p. 7.)

Lepore, Ernest

1986 (ed.), *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson*, Basil Blackwell, Oxford.

Lepore, Ernest and Kirk Ludwig

2006 (eds.), *The Essential Davidson*, Clarendon Press, Oxford.

Levy, Neil

2011 *Hard Luck: How Luck Undermines Free Wil and Moral Responsibility*, Oxford University Press, Oxford. (Cited on pp. 19, 30, 33, 36, 37, 51, 54.)

Lewis, David K.

1973 *Counterfactuals*, Blackwell, Oxford. (Cited on p. 182.)
1979 "Attitudes De Dicto and De Se," *Philosophical Review*, 88(4), pp. 513-543. (Cited on p. 113.)
1983 "New Work for a Theory of Universals," *Australasian Journal of Philosophy*, 61(December), pp. 343-377. (Cited on pp. 54, 182.)
1986 *On the Plurality of Worlds*, Blackwell, Oxford. (Cited on pp. 57, 185.)
1994 "Humean Supervenience Debugged," *Mind*, 103(412), pp. 473-490. (Cited on pp. 182, 185.)
1997 "Finkish Dispositions," *Philosophical Quarterly*, 47(187), pp. 143-158. (Cited on p. 188.)

Loewer, Barry

1996 "Humean Supervenience," *Philosophical Topics*, 24(1), pp. 101-127. (Cited on p. 182.)

Lowe, E. Jonathan

2008 *Personal Agency: The Metaphysics of Mind and Action*, Oxford University Press, USA. (Cited on pp. 35, 36, 183.)
2009 *More Kinds of Being: A Further Study of Individuation, Identity, and the Logic of Sortal Terms*, Wiley-Blackwell, Oxford. (Cited on p. 190.)
2013 "The Will as a Rational Free Power," in Groff *et al.* [2013], pp. 172-185. (Cited on pp. 182, 194.)

Maier, John

2013 "The Agentive Modalities," *Philosophy and Phenomenological Research*, 87(3), pp. 113-134. (Cited on p. 182.)

Malcolm, Norman

1968 "The Conceivability of Mechanism," *Philosophical Review*, 77(January), pp. 45-72. (Cited on pp. 175, 180.)

Markosian, Ned

1999 "A Compatibilist Version of the Theory of Agent Causation," *Pacific Philosophical Quarterly*, 80(3), pp. 257-277. (Cited on pp. 27, 42.)

Martin, C.B.

1994 "Dispositions and Conditionals," *Philosophical Quarterly*, 44(174), pp. 1-8. (Cited on p. 188.)

McDowell, John H.

1979 "Virtue and Reason," *The Monist*, 62(3), pp. 331-350. (Cited on p. 220.)
1981 "Non-Cognitivism & Rule-Following," in McDowell [1998], pp. 198-218. (Cited on p. 25.)

1994 *Mind and World*, Harvard University Press, Cambridge, MA. (Cited on pp. 4, 5, 61, 174.)
1998 *Mind, Value & Reality*, Harvard University Press, Cambridge, MA.
1999 *Reason and Nature*, ed. by Markus Willaschek, LIT Verlag, Münster. (Cited on pp. 5, 14, 227.)
2009 "Selections from Criteria, Defeasibility, and Knowledge," in Byrne *et al.* [2009], pp. 75-90. (Cited on p. 57.)

McGuire, J.M.
2007 "Actions, Reasons, and Intentions: Overcoming Davidson's Ontological Prejudice," *Dialogue*, 46, pp. 459-79. (Cited on p. 22.)

McKenna, Michael
2009 "Compatibilism," in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Winter 2009. (Cited on pp. 22, 26.)

McLaughlin, B.P. and J.D. Cohen
2007 (eds.), *Contemporary Debates in Philosophy of Mind*, Wiley-Blackwell, Oxford.

McLaughlin, Brian and Karen Bennett
2014 "Supervenience," in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Spring 2014. (Cited on p. 179.)

Melden, Abraham I.
1961 *Free Action*, Routledge & Kegan Paul, Abingdon. (Cited on pp. 173, 174, 176.)

Mele, Alfred R.
1992 *Springs of Action: Understanding Intentional Behavior*, Oxford University Press, New York. (Cited on p. 67.)
1997 (ed.), *The Philosophy of Action*, Oxford University Press, New York.
2006 *Free will and luck*, Oxford University Press, USA. (Cited on pp. 29, 30, 42, 43, 52, 70.)

Merricks, Trenton
2001 *Objects and Persons*, Oxford University Press, Oxford. (Cited on p. 195.)

Millikan, Ruth Garrett
1993 *White Queen Psychology and Other Essays for Alice*, The MIT Press, Cambridge, MA. (Cited on p. 22.)
1995 "Pushmi-Pullyu Representations," *Philosophical Perspectives*, 9, pp. 185-200. (Cited on p. 142.)

Molnar, George
2003 *Powers: A Study in Metaphysics*, ed. by Stephen Mumford, Oxford University Press, Oxford. (Cited on p. 187.)

Moran, Richard
2001 *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton University Press, Princeton. (Cited on pp. 77-80.)
2003 "Responses to O'Brien and Shoemaker," *European Journal of Philosophy*, 11(3), pp. 402-19. (Cited on p. 79.)
2004 "Anscombe on 'Practical Knowledge'," *Philosophy*, 55(2004), pp. 43-68. (Cited on p. 163.)

Moran, Richard and Martin J. Stone
2011a "Anscombe on Expression of Intention: An Exegisis," in Ford *et al.* [2011], pp. 198-210. (Cited on p. 127.)
2011b "Anscombe on the Expression of Intention: An Exegisis," in Ford *et al.* [2011], pp. 33-75. (Cited on p. 75.)

Morgan, Daniel
2015 "The Demonstrative Model of First-Person Thought," *Philosophical Studies*, 172(7), pp. 1795-1811. (Cited on pp. 74, 110.)

Mulder, Jesse M.
2016 "A Vital Challenge to Materialism," *Philsophy*, 91(2), pp. 153-182. (Cited on p. 197.)

Müller, Thomas
2012 "Branching in the Landscape of Possibilities," *Synthese*, 188(1), pp. 41-65. (Cited on p. 191.)

Mumford, Stephen
2003 *Powers: A Study in Metaphysics*, Clarendon Press, Oxford. (Cited on p. 187.)
Mumford, Stephen and Rani Lill Anjum
2011 *Getting Causes from Powers*, Oxford University Press, Oxford. (Cited on p. 187.)
2014 "A New Argument against Compatibilism," *Analysis*, 74(1), pp. 20-25. (Cited on pp. 182, 193.)
Nagel, Thomas
1986 *The View from Nowhere*, Oxford University Press, Oxford. (Cited on p. 31.)
O'Brien, Lucy
2007 *Self-Knowing Agents*, Oxford University Press, Oxford. (Cited on pp. 102, 107.)
O'Connor, Timothy
2002 *Persons and Causes: The Metaphysics of Free Will*, Oxford University Press, Oxford. (Cited on pp. 33, 68.)
2005 "Libertarian Views: Dualist and Agent-Causal Theories," in Kane [2005], ch. 15, pp. 337-356. (Cited on pp. 33-36.)
Papineau, David
1998 "Mind the Gap," *Philosophical Perspectives*, 12(12), pp. 373-89. (Cited on p. 198.)
Paul, Sarah K.
2009 "How We Know What We're Doing," *Philosophers' Imprint*, 9(11). (Cited on p. 141.)
Peels, Rik
2015 "A Modal Solution to the Problem of Moral Luck," *American Philosophical Quarterly*, 52(1). (Cited on p. 50.)
Pereboom, Derk
2002 "Living Without Free Will: The Case for Hard Incompatibilism," in Kane [2002], pp. 477-88. (Cited on pp. 34, 36.)
2003 "Source Incompatibilism and Alternative Possibilities," in Widerker *et al.* [2003], ch. 10, pp. 185-200. (Cited on p. 21.)
2007 "Hard Incompatibilism," in Fischer *et al.* [2007], pp. 85-125. (Cited on pp. 19, 31-33, 38, 200.)
Pickard, Hanna
2004 "Knowledge of action without observation," *Proceedings of the Aristotelian Society*, 104, pp. 205-230. (Cited on p. 125.)
Pritchard, Duncan
2005 *Epistemic Luck*, Clarendon Press. (Cited on pp. 50, 51, 54.)
2012 "Anti-Luck Virtue Epistemology," *The Journal of Philosophy*, 109(3), pp. 247-279. (Cited on pp. 55, 57.)
Pritchard, Duncan, Adrian Haddock, and Alan Millar
2010 *The Nature and Value of Knowledge: Three Investigations*, Oxford University Press, Oxford.
Rödl, Sebastian
2007 *Self-Consciousness*, Harvard University Press, Cambridge, MA. (Cited on pp. 7, 14, 56, 107-109, 113, 114, 116, 117, 138, 156, 159, 211, 213, 220, 224, 229.)
2012 *Categories of the Temporal: An Inquiry into the Forms of the Finite Understanding*, Harvard University Press, Cambridge, MA. (Cited on pp. 167, 188.)
Rosenthal, David
2006 *Consciousness and Mind*, Oxford University Press, Oxford. (Cited on p. 74.)
Russell, Bertrand
2009 *Human Knowledge: Its Scope and Limits*, Routledge, Abingdon. (Cited on p. 51.)
Ryle, Gilbert
1949/2009 *The Concept of Mind*, Routledge, New York. (Cited on pp. 175, 192, 234.)
Schlosser, Markus E.
2007 "Basic Deviance Reconsidered," *Analysis*, 67(295), pp. 186-194. (Cited on p. 154.)

Schwenkler, John
2011 "Perception and Practical Knowledge," *Philosophical Explorations*, 14(2), pp. 137-152. (Cited on p. 166.)
Searle, John R.
2003 *Rationality in Action*, The MIT Press, Cambridge, MA. (Cited on p. 24.)
Sehon, Scott
2012 "Action Explanation and the Free Will Debate: How Incompatibilist Arguments Go Wrong," *Philosophical Issues*, 22(1), pp. 351-368. (Cited on p. 176.)
Sellars, Wilfrid
2000 *Empiricism and the Philosophy of Mind*, ed. by Richard Rorty and Robert Brandom, Harvard University Press, Cambridge, MA. (Cited on pp. 56, 92.)
Setiya, Kieran
2007 *Reasons without Rationalism*, Princeton University Press, Princeton. (Cited on pp. 122, 141-146, 149.)
2008 "Practical Knowledge," *Ethics*, 118(3), pp. 388-409. (Cited on p. 144.)
2009 "Practical Knowledge Revisited," *Ethics*, 120(1), pp. 128-137. (Cited on p. 141.)
2011 "Knowledge of Intention," in Ford *et al.* [2011], pp. 170-197. (Cited on pp. 74, 75, 80-85, 91, 93, 102.)
2012 "Knowing How," *Proceedings of the Aristotelian Society*, 112(3), pp. 285-307. (Cited on pp. 144, 166.)
Shabo, Seth
2011 "Free Will and Mystery: Looking past the Mind Argument," *Philosophical Studies*, pp. 1-17. (Cited on pp. 19, 30, 40.)
2013 "Assimilations and Rollbacks: Two Arguments against Libertarianism Defended," *Philosophia*, pp. 1-22. (Cited on pp. 39, 44.)
Shoemaker, Sydney
1968 "Self-Reference and Self-Awareness," *Journal of Philosophy*, 65(19), pp. 555-67. (Cited on pp. 76, 104.)
Smith, Joel and Peter Sullivan
2011 (eds.), *Transcendental Philosophy and Naturalism*, Oxford University Press.
Smith, Michael
1987 "The Humean Theory of Motivation," *Mind*, 96(381). (Cited on p. 142.)
1995 *The Moral Problem*, Blackwell, Oxford. (Cited on pp. 23, 25.)
Sorell, Tom
2005 *Descartes Reinvented*, Cambridge University Press, Cambridge. (Cited on p. 110.)
Steward, Helen
2009 "Fairness, Agency and the Flicker of Freedom," *Nous*, 43(1), pp. 64-93. (Cited on p. 21.)
2012a *A Metaphysics for Freedom*, Oxford University Press. (Cited on pp. 4, 15, 26, 46, 172, 194, 202-206, 232.)
2012b "The Metaphysical Presuppositions of Moral Responsibility," *Journal of Ethics*, 16(2), pp. 241-271. (Cited on pp. 222, 223.)
Stoutland, Frederick
2009 "Von Wright's Compatibilism," in *Philosophical Probings. Essays on Von Wright's later work.* Ed. by Frederick Stoutland, Automatic Press, Copenhagen, pp. 61-82. (Cited on p. 175.)
Stroud, Barry
2011 *Engagement and Metaphysical Dissatisfaction: Modality and Value*, Oxford University Press, New York. (Cited on p. 66.)
Tanney, Julia
1995 "Why Reasons May Not be Causes," *Mind & Language*, 10(1-2), pp. 105-128. (Cited on p. 22.)
Tenenbaum, Sergio
2010 (ed.), *Desire, Practical Reason, and the Good*, Oxford University Press, New York.

Thompson, Michael

2004 "Apprehending Human Form," *Royal Institute of Philosophy Supplement*, 54, pp. 47-74. (Cited on pp. 208, 209.)

2008 *Life and Action: Elementary Structures of Practice and Practical Thought*, Harvard University Press, Cambridge, MA. (Cited on pp. 7, 113, 130, 157, 206, 207, 209, 210, 212, 214, 232.)

2011 "Anscombe's *Intention* and Practical Knowledge," in Ford *et al.* [2011], pp. 198-210. (Cited on pp. 6, 127, 160, 165, 166, 169.)

Van Inwagen, Peter

1986 *An Essay on Free Will*, Oxford University Press, Oxford. (Cited on pp. 19, 26-28, 183.)

2000 "Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture," *Philosophical Perspectives*, 14, pp. 1-19. (Cited on pp. 19, 39.)

Van Miltenburg, Niels

2015 *Freedom in Action*, PhD thesis, Utrecht University. (Cited on pp. 67, 150.)

Velleman, J. David

1989 *Practical Reflection*, Princeton University Press, Princeton. (Cited on pp. 122, 133, 136-138, 140.)

2000 *The Possibility of Practical Reason*, Oxford University Press. (Cited on pp. 137, 141.)

2005 "Précis of the Possibility of Practical Reason," *Philosophical Studies*, 121(3), pp. 225-238. (Cited on pp. 133, 134, 137.)

2006a *Self to Self: Selected Essays*, Cambridge University Press, Cambridge.

2006b "The Centered Self," in Velleman [2006a], pp. 253-283. (Cited on p. 133.)

Vihvelin, Kadri

2013 *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*, Oxford University Press, New York. (Cited on p. 182.)

Von Wright, Georg Henrik

1971 *Explanation and Understanding*, Cornell University Press, Ithaca, NY. (Cited on p. 176.)

1972 "On So-Called Practical Inference," *Acta Sociologica*, pp. 39-53. (Cited on pp. 175, 176.)

1974 *Causality and Determinism*, Columbia University Press, New York. (Cited on pp. 27, 225.)

Wedgwood, Ralph

2006 "The Normative Force of Reasoning," *Noûs*, 40(4), pp. 660-686. (Cited on p. 153.)

Widerker, David and Michael McKenna

2003 *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*, Ashgate, Aldershot. (Cited on p. 21.)

Wiggins, David

1987a *Needs, Values, Truth*, Blackwell, Oxford.

1987b "Towards a Reasonable Libertarianism," in Wiggins [1987a], pp. 269-302. (Cited on p. 44.)

2001 *Sameness and Substance Renewed*, Cambridge University Press, Cambridge. (Cited on p. 190.)

Wilson, George and Samuel Shpall

2012 "Action," in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Summer 2012. (Cited on p. 68.)

Wittgenstein, Ludwig

2001 [1953] *Philosophical Investigations*, translated by G.E.M. Anscombe., Wiley-Blackwell, Oxford. (Cited on pp. 23, 123, 124, 152.)

Yli-Vakkuri, Juhani

2010 "Conditional and Habitual Analyses of Disposition Ascriptions," *Philosophical Quarterly*, 60(240), pp. 624-630. (Cited on p. 188.)

Samenvatting in het Nederlands

Deze dissertatie onderzoekt een thema dat al eeuwenlang onderwerp is van wijsgerig debat: ons vermogen tot vrij handelen. Ondanks het klassieke academische karakter ervan, is dit een onderwerp waarmee we allemaal op een bijzondere manier vertrouwd zijn: we weten wat het is om vrij te handelen, omdat we vrij handelen. Dat is een bijzondere eigenschap van het vermogen tot handelen: we zijn niet met *alle* vermogens die we bezitten op dezelfde manier vertrouwd. Maar wanneer we een bepaalde handeling verrichten (bijvoorbeeld: ik beweeg mijn vingers over het toetsenbord om deze zin te typen) weten we dat wat er gebeurt (de woorden van deze zin verschijnen op het beeldscherm) *ons eigen werk* is. Dit is niet slechts een toevallige eigenschap van het vermogen tot vrij handelen: een gebeurtenis kan onmogelijk een handeling zijn—iets wat we met opzet, oftewel *intentioneel* doen—wanneer we die handeling niet als zodanig beschouwen. Als we niet op deze bijzondere manier vertrouwd waren met onze handelingen, zouden er, zo lijkt het, helemaal geen vrije handelingen zijn. Een onderzoek naar ons vermogen tot vrij handelen moet dan ook noodzakelijkerwijs een onderzoek zijn naar onze vertrouwdheid met—onze *kennis* van—dit vermogen. Het moet een onderzoek zijn naar ons zelfbegrip als vrije wezens, dat de basis vormt van een groot deel van ons doen en denken. Dat verklaart de titel van deze dissertatie: vrijheid en zelfkennis zijn geen gescheiden onderwerpen.

Inleiding

Onze vertrouwdheid met ons vermogen tot vrij handelen wordt geproblematiseerd wanneer we filosofisch en wetenschappelijk reflecteren op onze plek in de natuurlijke orde. Het lijkt er immers op, zo wordt vaak beweerd, dat de wetenschap ons leert dat alles wat er gebeurt een schakel is in een lange ketting van oorzaak en gevolg. Wanneer we een gebeurtenis op die manier beschouwen, als het noodzakelijke gevolg van de stand van zaken aan het begin der tijden, dan lijken we daarmee te zeggen dat deze gebeurtenis juist *niet* aan ons toe te schrijven valt, en dus juist *geen* handeling

is. Er bestaat dus een spanning tussen het begrip van een handeling en de hypothese dat alles *gedetermineerd* is—een hypothese die, paradoxaal genoeg, zelf het resultaat is van het verrichten van wetenschappelijke handelingen.

In de wijsbegeerte bestaan er grofweg twee scholen van denken over deze problematiek: het zogenaamde *compatibalisme* en *incompatibilisme*. De incompatibilist verdedigt de bovenstaande redenering: wanneer alles gedetermineerd zou zijn, zou vrij handelen onmogelijk zijn. De compatibilist beweert daarentegen juist dat de beschreven spanning tussen vrijheid en determinisme slechts schijn is: vrij handelen betekent eenvoudigweg niets anders dan dat wat men doet veroorzaakt wordt door bepaalde psychologische gebeurtenissen, bijv. 'verlangens' of 'intenties', die aan de handeling voorafgaan (daaraan wordt vaak toegevoegd dat deze psychologische gebeurtenissen terug te vinden zijn in de hersenen).

Nu wijst de incompatibilist er vaak (terecht) op dat volgens de moderne fysica het determinisme helemaal niet waar is. Op het niveau van de allerkleinste deeltjes volgt niet elke gebeurtenis noodzakelijk op de voorgaande. Vaak werd eenvoudigweg aangenomen dat het bestaan van zulke fysische onbepaaldheid (indeterminisme) het bestaan van de vrije wil kan redden. Maar steeds vaker wijzen filosofen erop dat ook indeterminisme op gespannen voet staat met het idee van vrij handelen. Wanneer een gebeurtenis (bijv. een menselijke armbeweging) niet bepaald zou worden door een voorgaande gebeurtenis, zo is de gedachte, dan heerst er chaos: alles wat er gebeurt is een kwestie van *toeval*. Het begrip van vrij handelen wordt dus bedreigd van beide kanten: de mogelijkheid om *zelf* te bepalen wat je doet verdwijnt, wanneer we alle gebeurtenissen beschouwen als noodzakelijk gevolg van een vorige gebeurtenis, maar ook als we alle gebeurtenissen beschouwen als het resultaat van een kosmische dobbelsteenworp. Beide gevallen lijken de mogelijkheid om onszelf als handelende wezens te beschouwen te ondermijnen.

Het doel van dit proefschrift is om het hierboven geschetste dilemma te ontrafelen, en te laten zien dat ons vermogen tot vrij handelen een vermogen is voor daadwerkelijke zelfbepaling: een intentionele handeling is noch een kwestie van toeval, noch het noodzakelijk gevolg van eerdere (hersen)toestanden en gebeurtenissen. Daarmee wordt duidelijk dat ons zelfbegrip als handelende wezens niet op gespannen voet hoeft te staan met een wetenschappelijke beschouwing van onszelf.

Vrijheid, determinisme, en toeval

In **hoofdstuk 1** bespreek ik dit dilemma tussen determinisme en toeval in de context van het moderne debat in de analytische traditie van de wijsbegeerte. Ik laat zien dat dit debat gestructureerd wordt door twee gerelateerde aannamen, die in de literatuur

helaas zelden expliciet gemaakt worden, en die ten grondslag liggen aan het dilemma. De eerste aanname betreft de aard van het intentioneel handelen. Bijna alle filosofen in het vrije wil debat hangen de zogenaamde causale handelingstheorie (*Causal Theory of Action*, of CTA) aan. Deze theorie beweert dat een intentionele handeling een gebeurtenis is die veroorzaakt wordt door bepaalde mentale gebeurtenissen: A doet intentioneel φ dan en slechts dan als φ een gebeurtenis is die veroorzaakt wordt door een *reden* van A. Redenen zijn volgens deze theorie mentale gebeurtenissen die op de een of andere manier terug te vinden moeten zijn, of identiek zijn aan, gebeurtenissen in de hersenen. Het betreft een *reductieve* theorie van handelen, omdat wordt aangenomen dat het soort oorzakelijkheid dat reden en handeling verbindt dezelfde is als die elders in de natuur voorkomt. CTA is een theorie die dus goed past bij compatibilistische opvattingen over vrije wil. Maar vreemd genoeg hangen ook vrijwel alle incompatibilisten in het huidige debat deze theorie aan. Ik laat zien dat het deze dominantie van de CTA is die ervoor zorgt dat incompatibilistische theorieën van vrije wil steeds bezwijken aan het toevalsprobleem (*luck objection*).

De tweede aanname die het vrije wil debat structureert is er een over de aard van vrijheid. Aangenomen wordt dat vrijheid een begrip is, en handelen een ander. De vraag van het vrije wil debat verschijnt dan als volgt: onder welke condities is een handeling vrij? Dat veronderstelt dat we al begrijpen wat een handeling is (want dat, zo wordt aangenomen, vertelt CTA ons), en dat we los daarvan begrijpen wat vrijheid is. Dat is een dogmatische aanname, die grotendeels nieuw is in de geschiedenis van de wijsbegeerte (alhoewel ze voorlopers kent in bijvoorbeeld Locke). Toch kan deze aanname onontkoombaar lijken: hoe anders kunnen we 'vrij' in 'vrije wil' of 'vrije handeling' begrijpen, dan als een bijvoegelijk naamwoord dat een losstaand begrip ('wil', 'handeling') nader specificeert?

Aan het einde van hoofdstuk 1 suggereer ik dat er een alternatief bestaat: vrijheid kunnen we ook beschouwen als een intrinsieke eigenschap van (intentionele) handelingen, op zo'n manier dat de twee onlosmakelijk verbonden zijn. We begrijpen niet wat een handeling is, als we niet begrijpen wat vrijheid is—en andersom. Zo wordt de vrije wil in een groot deel van de wijsgerige traditie begrepen. Maar, zo stel ik, de dominantie van de CTA maakt het onmogelijk om te zien hoe vrijheid een intrinsieke eigenschap van de wil (het vermogen tot handelen) kan zijn. Als de CTA klopt, dan is immers elke handeling of gedetermineerd, of een kwestie van toeval. Zolang we vasthouden aan deze twee aannames, blijft het onmogelijk om aan het dilemma te ontkomen.

Om de weg vrij te maken naar een alternatief begrip van vrij handelen, bespreek ik in **hoofdstuk 2** in meer detail de vraag waarom toeval en vrijheid elkaar lijken uit te sluiten. Ik laat zien dat er hiervoor binnen de termen van het huidige debat eigenlijk

geen goede onderbouwing gegeven kan worden. Dat komt omdat er gewerkt wordt met een *uniforme notie van toeval*: het idee dat er maar een soort noodzakelijkheid bestaat, en dat alles wat niet in die zin noodzakelijk (gedetermineerd) is, toevallig is. Daartegen beargumenteer ik dat er verschillende vormen van noodzakelijkheid zijn, en er evenzoveel corresponderende noties van toeval zijn. Een gebeurtenis die niet vooraf gedetermineerd is, is in bepaalde zin contingent of toevallig—maar niet noodzakelijk in dezelfde betekenis van 'toevallig' die vrijheid ondermijnt. Dat komt, zo stel ik, omdat vrijheid zelf een *vorm van bepaling*—en dat wil zeggen een vorm van *causaliteit*—is. De CTA neemt dogmatisch aan dat er maar een vorm van causaliteit, en dus maar een vorm van bepaling is. Ik laat zien dat de standaardargumentatie voor CTA (naar Donald Davidson) daarom niet klopt.

Als een gebeurtenis een vrije handeling is, dan is ze niet toevallig maar noodzakelijk. Maar het is niet duidelijk of deze vorm van noodzakelijkheid (vrijheid) compatibel is met, of zelfs vereist dat, een handeling ook ook noodzakelijk is in de zin van determinisme. Ik beargumenteer dat we om die vraag—de vraag naar de waarheid van compatibilisme of incompatibilisme—te kunnen beantwoorden eerst een alternatief voor de CTA moeten ontwikkelen.

Zelfkennis en causaliteit

De alternatieve, niet-reductieve theorie van intentioneel handelen die ik wil ontwikkelen volgt de inzichten van de Britse filosofe Elizabeth Anscombe. Volgens Anscombe wordt de vorm van bepaling die ten grondslag ligt aan intentioneel handelen gekenmerkt door *praktische kennis*: kennis die we hebben van wat we aan het doen zijn. Deze kennis is niet gebaseerd op observatie: we hoeven bijvoorbeeld niet van het feit dat we op de Voorstraat aan het lopen zijn af te leiden dat we op weg zijn naar de supermarkt—zo'n afleiding zou onmogelijk zijn, en ze is bovendien ook niet nodig. We weten eenvoudigweg of we op weg zijn naar de supermarkt, of naar het café. Praktische kennis is een vorm van zelfkennis. In de analytische wijsbegeerte wordt dit vaak als volgt omschreven: we weten *vanuit de eerste persoon* wat we aan het doen zijn.

In de contemporaine filosofie wordt zelfkennis echter begrepen op een manier die even reductief is als de CTA. Voor we dus een adequate notie van praktische kennis kunnen ontwikkelen, die op haar beurt een goed begrip van vrijheid mogelijk maakt, moeten we dit reductieve begrip van zelfkennis weerleggen. Dat doe ik in **hoofdstuk 3**. Alvorens over te gaan naar praktische zelfkennis—kennis van onze intenties en onze handelingen—bespreek ik daar eerst het geval van theoretische zelfkennis: kennis van onze *overtuigingen*.

Volgens de heersende aannames in de filosofie van de geest is een overtuiging een zogenaamde propositionele attitude: een houding die een subject inneemt ten opzichte van een propositie zoals 'de kat zit op de mat'. Iemand die gelooft dat de kat op de mat zit, neemt de houding 'overtuigd-zijn-dat' tegenover die propositie in. Wanneer ik nu reflecteer op het feit dat ik geloof dat de kat op de mat zit, kan ik komen tot een *tweede-orde overtuiging*: de overtuiging 'ik geloof dat de kat op de mat zit'. Ik kan echter ook de eerste-orde overtuiging bezitten, zonder de tweede-orde overtuiging. Ik laat zien dat dit begrip van zelfkennis en propositionele attitudes uiterst problematisch is: als zelfkennis een kwestie zou zijn van het vormen van tweede-orde overtuigingen, dan zouden we nooit tot *kennis* van onze eigen overtuigingen kunnen komen. Wat legitimeert immers de overgang van een overtuiging 'p' naar de overtuiging 'ik geloof dat p'?

In plaats daarvan beargumenteer ik dat zelfkennis en overtuiging veel nauwer verbonden zijn. Wanneer ik tot de conclusie kom dat p, oordeel ik dat het juist is om p te geloven. Maar dit oordeel, dat ik p moet geloven omdat ik bijvoorbeeld al geloof dat q waar is, *is zelf* het overtuigd-zijn-dat-p. Wanneer ik overtuigd ben dat p, bevat deze overtuiging dus als het ware al het gegeven *dat ik het geloof*. Er bestaat dus niet zoiets als een overtuiging die niet (tenminste impliciet) binnen de grenzen van het zelfbewustzijn valt. Ik volg hier het werk van onder andere Sebastian Rödl, die deze gedachte uit de wijsgerige traditie in de analytische context weer heeft ingevoerd.

Er bestaat dus een nauw verband tussen zelfkennis en zelfbepaling. Te weten komen dat ik geloof dat p is niets anders dan concluderen dat p waar is: het opdoen van zelfkennis is het bepalen wat ik denk. Op basis hiervan laat ik zien dat *redeneren*—het soort oorzakelijkheid waarmee we te maken hebben als ik denk 'ik geloof dat p *omdat* ik geloof dat q'—niet de mechanistische, reductieve notie van causaliteit kan zijn die ten grondslag ligt aan de CTA. Het is een vorm van causaliteit die intrinsiek gekenmerkt wordt door zelfbewustzijn. De traditionele naam voor deze vorm van causaliteit is *spontaniteit*.

In **hoofdstuk 4** zet ik dit begrip van zelfkennis en spontaniteit in in de context van het handelen. Ik beargumenteer dat we praktische kennis ook niet kunnen begrijpen volgens het tweede-orde model (namelijk als een overtuiging 'ik doe φ'). Het is onmogelijk om intentioneel te handelen zonder praktische kennis. Ik laat zien hoe hier definitief uit blijkt dat de CTA een onware theorie over handelen is. Praktische kennis is ook spontane zelfkennis. Door praktisch te redeneren over wat ik *moet* doen bepaal ik wat ik daadwerkelijk *aan het doen ben*: praktische kennis is het antwoord op de vraag 'wat moet ik doen?', en *is* daarom de handeling zelf.

Ook praktisch redeneren is dus een zelfbewuste, spontane activiteit. Maar laat dat ook zien dat we, wanneer we intentioneel handelen, vrij zijn in de zin dat onze

handeling niet door eerdere gebeurtenissen gedetermineerd is? In **hoodstuk 5** beargumenteer ik dat dit inderdaad het geval is: de spontaniteit van praktische kennis impliceert indeterminisme.

Om dit te beargumenteren ontwikkel ik een omvattend raamwerk van verschillende soorten causaliteit. Ik betoog dat causaliteit in al haar vormen een manifestatie is van de *vermogens* van objecten of organismen. Ik maak een systematisch onderscheid tussen verschillende soorten vermogens: de vermogens van objecten in de levenloze natuur, de vermogens van organismen, en de vermogens van zelfbewuste subjecten zoals wij. Omdat er in de levenloze natuur substanties zijn die indeterministische vermogens hebben (vermogens die zich niet noodzakelijkerwijs manifesteren), is het mogelijk dat er complexere organismen bestaan met het vermogen om intentioneel te handelen.

Ik laat zien dat uit de zelfbewuste aard van intentioneel handelen volgt dat dit vermogen zelf ook indeterministisch moet zijn. Het vermogen om intentioneel te handelen manifesteert zich immers wanneer het subject oordeelt dat het juist is om de handeling te verrichten. Maar aangezien dit oordeel niets anders is dan de handeling zelf, volgt het dat het vermogen tot handelen *zelf* bepaalt of dat vermogen zich al dan niet manifesteert. Een intentionele handeling is daarom een vooraf onbepaalde gebeurtenis, die echter geen kwestie van louter toeval is: het vermogen tot handelen is een daadwerkelijk *vrij* vermogen.

* * *

Curriculum Vitae

Dawa Ometto was born on November 3rd 1986 in De Bilt, the Netherlands. He graduated with bachelor degrees in philosophy and in history from Utrecht University in 2009. He obtained a master degree (*cum laude*) in philosophy at Utrecht University in 2012, after which he was employed at the same university as a PhD-researcher. In 2016 he was a visiting scholar at the University of Leipzig for the duration of one semester. He has given courses, published articles, and given talks on topics in the philosophy of action, ethics, philosophy of science, and philosophy of mind. He happily lives in Utrecht with his partner Candice.

* * *

Quaestiones Infinitae

PUBLICATIONS OF THE DEPARTMENT OF PHILOSOPHY AND RELIGIOUS STUDIES

VOLUME 21. D. VAN DALEN, *Torens en Fundamenten* (valedictory lecture), 1997.
VOLUME 22. J.A. BERGSTRA, W.J. FOKKINK, W.M.T. MENNEN, S.F.M. VAN VLIJMEN, *Spoorweglogica via EURIS*, 1997.
VOLUME 23. I.M. CROESE, *Simplicius on Continuous and Instantaneous Change* (dissertation), 1998.
VOLUME 24. M.J. HOLLENBERG, *Logic and Bisimulation* (dissertation), 1998.
VOLUME 25. C.H. LEIJENHORST, *Hobbes and the Aristotelians* (dissertation), 1998.
VOLUME 26. S.F.M. VAN VLIJMEN, *Algebraic Specification in Action* (dissertation), 1998.
VOLUME 27. M.F. VERWEIJ, *Preventive Medicine Between Obligation and Aspiration* (dissertation), 1998.
VOLUME 28. J.A. BERGSTRA, S.F.M. VAN VLIJMEN, *Theoretische Software-Engineering: kenmerken, faseringen en classificaties*, 1998.
VOLUME 29. A.G. WOUTERS, *Explanation Without A Cause* (dissertation), 1999.
VOLUME 30. M.M.S.K. SIE, *Responsibility, Blameworthy Action & Normative Disagreements* (dissertation), 1999.
VOLUME 31. M.S.P.R. VAN ATTEN, *Phenomenology of choice sequences* (dissertation), 1999.
VOLUME 32. V.N. STEBLETSOVA, *Algebras, Relations and Geometries (an equational perspective)* (dissertation), 2000.
VOLUME 33. A. VISSER, *Het Tekst Continuüm* (inaugural lecture), 2000.
VOLUME 34. H. ISHIGURO, *Can we speak about what cannot be said?* (public lecture), 2000.
VOLUME 35. W. HAAS, *Haltlosigkeit; Zwischen Sprache und Erfahrung* (dissertation), 2001.
VOLUME 36. R. POLI, *ALWIS: Ontology for knowledge engineers* (dissertation), 2001.
VOLUME 37. J. MANSFELD, *Platonische Briefschrijverij* (valedictory lecture), 2001.
VOLUME 37A. E.J. BOS, *The Correspondence between Descartes and Henricus Regius* (dissertation), 2002.
VOLUME 38. M. VAN OTEGEM, *A Bibliography of the Works of Descartes (1637-1704)* (dissertation), 2002.
VOLUME 39. B.E.K.J. GOOSSENS, *Edmund Husserl: Einleitung in die Philosophie: Vorlesungen 1922/23* (dissertation), 2003.
VOLUME 40. H.J.M. BROEKHUIJSE, *Het einde van de sociaaldemocratie* (dissertation), 2002.
VOLUME 41. P. RAVALLI, *Husserls Phänomenologie der Intersubjektivität in den Göttinger Jahren: Eine kritisch-historische Darstellung* (dissertation), 2003.
VOLUME 42. B. ALMOND, *The Midas Touch: Ethics, Science and our Human Future* (inaugural lecture), 2003.
VOLUME 43. M. DÜWELL, *Morele kennis: over de mogelijkheden van toegepaste ethiek*

(inaugural lecture), 2003.

VOLUME 44. R.D.A. HENDRIKS, *Metamathematics in Coq* (dissertation), 2003.

VOLUME 45. TH. VERBEEK, E.J. BOS, J.M.M. VAN DE VEN, *The Correspondence of René Descartes: 1643*, 2003.

VOLUME 46. J.J.C. KUIPER, *Ideas and Explorations: Brouwer's Road to Intuitionism* (dissertation), 2004.

VOLUME 47. C.M. BEKKER, *Rechtvaardigheid, Onpartijdigheid, Gender en Sociale Diversiteit; Feministische filosofen over recht doen aan vrouwen en hun onderlinge verschillen* (dissertation), 2004.

VOLUME 48. A.A. LONG, *Epictetus on understanding and managing emotions* (public lecture), 2004.

VOLUME 49. J.J. JOOSTEN, *Interpretability formalized* (dissertation), 2004.

VOLUME 50. J.G. SIJMONS, *Phänomenologie und Idealismus: Analyse der Struktur und Methode der Philosophie Rudolf Steiners* (dissertation), 2005.

VOLUME 51. J.H. HOOGSTAD, *Time tracks* (dissertation), 2005.

VOLUME 52. M.A. VAN DEN HOVEN, *A Claim for Reasonable Morality* (dissertation), 2006.

VOLUME 53. C. VERMEULEN, *René Descartes, Specimina philosophiae: Introduction and Critical Edition* (dissertation), 2007.

VOLUME 54. R.G. MILLIKAN, *Learning Language without having a theory of mind* (inaugural lecture), 2007.

VOLUME 55. R.J.G. CLAASSEN, *The Market's Place in the Provision of Goods* (dissertation), 2008.

VOLUME 56. H.J.S. BRUGGINK, *Equivalence of Reductions in Higher-Order Rewriting* (dissertation), 2008.

VOLUME 57. A. KALIS, *Failures of agency* (dissertation), 2009.

VOLUME 58. S. GRAUMANN, *Assistierte Freiheit* (dissertation), 2009.

VOLUME 59. M. AALDERINK, *Philosophy, Scientific Knowledge, and Concept Formation in Geulincx and Descartes* (dissertation), 2010.

VOLUME 60. I.M. CONRADIE, *Seneca in his cultural and literary context: Selected moral letters on the body* (dissertation), 2010.

VOLUME 61. C. VAN SIJL, *Stoic Philosophy and the Exegesis of Myth* (dissertation), 2010.

VOLUME 62. J.M.I.M. LEO, *The Logical Structure of Relations* (dissertation), 2010.

VOLUME 63 M.S.A. VAN HOUTE, *Seneca's theology in its philosophical context* (dissertation), 2010.

VOLUME 64 F.A. BAKKER, *Three Studies in Epicurean Cosmology* (dissertation), 2010.

VOLUME 65 T. FOSSEN, *Political legitimacy and the pragmatic turn* (dissertation), 2011.

VOLUME 66 T. VISAK, *Killing happy animals. Explorations in utilitarian ethics.* (dissertation), 2011.

VOLUME 67 A. JOOSSE, *Why we need others: Platonic and Stoic models of friendship and self-understanding* (dissertation), 2011.

VOLUME 68 N. M. NIJSINGH, *Expanding newborn screening programmes and strengthening informed consent* (dissertation), 2012.

VOLUME 69 R. PEELS, *Believing Responsibly: Intellectual Obligations and Doxastic Excuses* (dissertation), 2012.
VOLUME 70 S. LUTZ, *Criteria of Empirical Significance* (dissertation), 2012
VOLUME 70A G.H. BOS, *Agential Self-consciousness, beyond conscious agency* (dissertation), 2013.
VOLUME 71 F.E. KALDEWAIJ, *The animal in morality: Justifying duties to animals in Kantian moral philosophy* (dissertation), 2013.
VOLUME 72 R.O. BUNING, *Henricus Reneri (1593-1639): Descartes' Quartermaster in Aristotelian Territory* (dissertation), 2013.
VOLUME 73 I.S. LÖWISCH, *Genealogy Composition in Response to Trauma: Gender and Memory in 1 Chronicles 1-9 and the Documentary Film 'My Life Part 2'* (dissertation), 2013.
VOLUME 74 A. EL KHAIRAT, *Contesting Boundaries: Satire in Contemporary Morocco* (dissertation), 2013.
VOLUME 75 A. KROM, *Not to be sneezed at. On the possibility of justifying infectious disease control by appealing to a mid-level harm principle* (dissertation), 2014.
VOLUME 76 Z. PALL, *Salafism in Lebanon: local and transnational resources* (dissertation), 2014.
VOLUME 77 D. WAHID, *Nurturing the Salafi Manhaj: A Study of Salafi Pesantrens in Contemporary Indonesia* (dissertation), 2014.
VOLUME 78 B.W.P VAN DEN BERG, *Speelruimte voor dialoog en verbeelding. Basisschoolleerlingen maken kennis met religieuze verhalen* (dissertation), 2014.
VOLUME 79 J.T. BERGHUIJS, *New Spirituality and Social Engagement* (dissertation), 2014.
VOLUME 80 A. WETTER, *Judging By Her. Reconfiguring Israel in Ruth, Esther and Judith* (dissertation), 2014.
VOLUME 81 J.M. MULDER, *Conceptual Realism. The Structure of Metaphysical Thought* (dissertation), 2014.
VOLUME 82 L.W.C. VAN LIT, *Eschatology and the World of Image in Suhrawardī and His Commentators* (dissertation), 2014.
VOLUME 83 P.L. LAMBERTZ, *Divisive matters. Aesthetic difference and authority in a Congolese spiritual movement 'from Japan'* (dissertation), 2015.
VOLUME 84 J.P. GOUDSMIT, *Intuitionistic Rules: Admissible Rules of Intermediate Logics* (dissertation), 2015.
VOLUME 85 E.T. FEIKEMA, *Still not at Ease: Corruption and Conflict of Interest in Hybrid Political Orders* (dissertation), 2015.
VOLUME 86 N. VAN MILTENBURG, *Freedom in Action* (dissertation), 2015.
VOLUME 86A P. COPPENS, *Seeing God in This world and the Otherworld: Crossing Boundaries in Sufi Commentaries on the Qurʾān* (dissertation), 2015.
VOLUME 87 D.H.J. JETHRO, *Aesthetics of Power: Heritage Formation and the Senses in Post-apartheid South Africa* (dissertation), 2015.

VOLUME 88 C.E. HARNACKE, *From Human Nature to Moral Judgement: Reframing Debates about Disability and Enhancement* (dissertation), 2015.

VOLUME 89 X. WANG, *Human Rights and Internet Access: A Philosophical Investigation* (dissertation), 2016.

VOLUME 90 R. VAN BROEKHOVEN, *De Bewakers Bewaakt: Journalistiek en leiderschap in een gemediatiseerde democratie* (dissertation), 2016.

VOLUME 91 A. SCHLATMANN, *Shi'i Muslim youth in the Netherlands: Negotiating Shi'i fatwas and rituals in the Dutch context* (dissertation), 2016.

VOLUME 92 M.L. VAN WIJNGAARDEN, *Schitterende getuigen. Nederlands luthers avondmaalsgerei als indenteitsdrager van een godsdienstige minderheid* (dissertation), 2016.

VOLUME 93 S. COENRADIE, *Vicarious substitution in the literary work of Shūsaku Endō. On fools, animals, objects and doubles* (dissertation), 2016.

VOLUME 94 J. RAJAIAH, *Dalit Humanization. A quest based on M.M. Thomas' theology of salvation and humanization* (dissertation), 2016.

VOLUME 95 D.L.A. OMETTO, *Freedom & Self-knowledge* (dissertation), 2016.