The Metaphysics of Agency

Markus Ernst Schlosser

Submitted for the Degree of Ph.D. in Philosophy

August 21, 2006

Abstract

Mainstream philosophy of action and mind construes intentional behaviour in terms of causal processes that lead from agent-involving mental states to action. Actions are construed as events, which are actions in virtue of being caused by the right mental antecedents in the right way. Opponents of this standard event-causal approach have criticised the view on various grounds; they argue that it does not account for free will and moral responsibility, that it does not account for action done in the light of reasons, or, even, that it cannot capture the very phenomenon of agency. The thesis defends the standard event-causal approach against challenges of that kind.

In the first chapter I consider theories that stipulate an irreducible metaphysical relation between the *agent* (or the *self*) and the action. I argue that such theories do not add anything to our understanding of human agency, and that we have, therefore, no reason to share the metaphysically problematic assumptions on which those alternative models are based. In the second chapter I argue for the claim that reason-explanations of actions are causal explanations, and I argue against non-causal alternatives. My main point is that the causal approach is to be preferred, because it provides an integrated account of agency by providing an account of the relation between the causes of movements and reasons for actions. In the third chapter I defend non-reductive physicalism as the most plausible version of the standard event-causal theory. In the fourth and last chapter I argue against the charge that the standard approach cannot account for the *agent's* role in the performance of action. Further, I propose the following stance with respect to the problem of free will: we do not have free will, but we have the related ability to govern ourselves—and the best account of self-determination presupposes causation, but not causal determinism.

Declarations

I, Markus Schlosser, hereby certify that this thesis, which is approximately 98.000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

Date: 30/02/07 Signature of candidate: Markus Schlosser

I was admitted as a research student in September, 2002 and as a candidate for the degree of Ph.D. in philosophy in September, 2003; the higher study for which this is a record was carried out in the University of St. Andrews between 2002 and 2006.

Date: 30/02/07 Signature of candidate: Markus Schlosser

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Ph.D. in philosophy in the University of St. Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 30/02/07 Signature of supervisor: Sarah Broadie

In submitting this thesis to the University of St. Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker.

Date: 30/02/07 Signature of candidate: Markus Schlosser

Acknowledgements

I would like to thank Sarah Broadie, Allan Millar, Robert Kane, Alfred Mele, Jonathan Lowe, John Haldane, Simon Prosser, James Harris, Peter Clarke, David Archard, Zoë Payne, and audiences at seminars and conferences at the Universities of St. Andrews, Madrid, Austin, Glasgow, Edinburgh and Hertfordshire for their helpful comments and replies. I am especially grateful to my supervisor Sarah Broadie for her comments, advice, and encouragement.

Research for this thesis was funded by a Postgraduate Research Grant from the Austrian Ministry of Education (2002-2004), an AHRC Postgraduate Award (2003-2005), and a S.A.S.P. Scholarship from the University of St. Andrews (2005-2006).

A slightly different version of the first part of the section 'The Causal Exclusion Argument' (Chapter 3) was first published as 'Causal Exclusion and Over-determination', in Di Nucci & McHugh (eds.) *Content, Consciousness and Perception: Essays in Contemporary Philosophy of Mind*, London: Cambridge Scholars Press, 2006.

Contents

Introduction

This thesis defends what I call the standard-causal model of agency. In broad outline, this theory construes intentional, rational and autonomous agency in terms of event-causal processes that lead from the agent's mental states and events to the performance of an action. It assumes that actions are events, which are actions in virtue of being caused by the right mental antecedents in the right way. The standard-causal approach has a long history.[1] But I shall restrict my considerations to recent and contemporary analytical philosophy. In the nineteen fifties and sixties, the theory was strongly criticised by many philosophers who were influenced by Wittgenstein's later philosophy and Ryle's *The Concept of Mind*.[2] It is generally agreed, though, that Donald Davidson successfully rebutted all the major objections in his seminal paper 'Actions, Reasons, and Causes'.[3]

Why, then, is the standard-causal model in need of defence? There are several things to say. The standard-causal model has been the predominant position in the analytical philosophy of action and mind, which is to say that it is largely taken for granted. However, throughout the history of philosophy one can find voices of dissent.[4] Given that, it is certainly interesting and worthwhile to question reigning orthodoxy and to take a new look at the assumptions behind it and at the intuitions that support it. However, the voices of dissent are not confined to times pre-analytical philosophy or pre-Davidson. In fact, in recent and contemporary analytical philosophy there is a trend towards alternative models of agency. I think we can distinguish between the following three main reasons responsible for that trend.

---

[1] Alfred Mele claims that the idea that an agent's mental states and events cause and causally explain actions is 'at least as old as Aristotle': 'the origin of action—its efficient not its final cause—is choice, and choice is desire and reasoning with a view to an end' (*Nichomachean Ethics*, 1139a31-32). Compare Mele, 2003, p. 38.

[2] Compare Anscombe, 1957; Melden, 1961; and Taylor, 1966.

[3] Davidson, 1963, reprinted in Davidson 1980, essay 1. G. F. Schueler, for instance, who is a contemporary opponent of the standard-causal theory, acknowledges that Davidson 'demolished' the early challenges to the standard-causal theory—especially the so-called logical connection argument. Compare Schueler, 2003, p. 9.

[4] There are, of course, numerous examples of philosophers who reject the approach to construe action in terms of efficient- or event-causation for various reasons. Many contemporary opponents of the view refer to Kant or Thomas Reid—or to Roderick Chisholm as an early opponent in the history of analytical philosophy. And, ironically, opponents of the view may refer to Aristotle, just as its proponents. In the *Physics*, for instance, Aristotle says that 'a staff moves a stone, and is moved by a hand, which is moved by a man' (VII, 5, 256a, 6-8)—as opposed, of course, to being moved by some state of or change in the man.

Firstly, precise and compelling formulations of the so-called consequence argument for incompatibilism about free will gained more and more weight in the recent debate on free will.[5] There is no straightforward connection to the question of whether the standard-causal model of agency is true or tenable. But it is not very difficult to see—as I will explain in chapters one and four—that the standard-causal model cannot account for libertarian free will—that is, the variety of free will that is incompatible with the thesis of causal determinism. Given that, it is only understandable that incompatibilists have been trying to find an alternative model of agency.

Secondly, the standard-causal model appears counterintuitive in at least the following two respects. The theory construes agency in terms of event-causal processes: actions are caused by mental states and events. Some philosophers think that this approach fails to capture the very phenomenon of agency, as it recognises only happenings and causal relations between them. When we act for reasons, for instance, we act spontaneously in the light of reasons. If that agential power is reduced to the causal efficacy of mental states and events that represent our reasons, then agency, as the challenge goes, *disappears*. Further, the standard-causal theory claims that reason-explanations of action are causal explanations. However, the practice of explaining human actions in terms of reasons is very different from the scientific practice of uncovering lawful connections between types of events.

Thirdly, some philosophers are dissatisfied with the reductive and non-reductive varieties of physicalism that have dominated recent philosophy of mind. Those theories of the mind typically presuppose the standard-causal model, which is a reductive model in the sense that it reduces the agent's power to act to the causal efficacy of agent-involving mental states and events. Alternatively, some philosophers seek to understand the relation between mind and body and the phenomenon of mental causation in terms of emergence and emergent causal powers. And some forms of emergentism, as I will explain in the first chapter, go hand in hand with non-reductive theories of agency.

Given all that, we can see why the standard-causal model is in need of defence. I will respond to all three challenges in due course. In the first chapter, I will defend the

[5] The most prominent statement of that argument can be found in van Inwagen, 1983. For an overview of the contemporary free will debate see Kane, 2002.

theory against the claim that non-reductive models of agency—in particular, agent-causal models—do better in capturing certain important features of human agency, such as free will and moral responsibility.

In the second chapter, I will defend causalism about reason-explanation—that is, the view that reasons cause and causally explain actions, which is an integral part of the standard-causal model. I will argue that non-causal alternatives fail provide an account of what it is to act *for* reasons, and I will provide what can be called a *global* argument for causalism. I will argue that causalism is to be preferred, because it is the only model that provides an *integrated* account of agency: it locates the phenomenon of action within the event-causal order and provides, thereby, an account of the relationship between reasons for actions and the causes of bodily movements.

In the third chapter, I will defend causalism in the form of non-reductive physicalism against the influential causal exclusion argument and against, what I shall call, the coincidence problem—the problem of explaining the systematic relationship between the level of intentional reason-explanation and lower levels of neuro-physiological and physical explanation. I will argue that this problem can be solved without assuming identities between mental and non-mental types and without assuming that intentional theories can be reduced.

In the fourth and in final chapter, I will respond to the mentioned challenge of disappearing agency. I will argue that the standard-causal model captures agency, because it captures the phenomenon of agential control. As part of that response, I will present a solution to the much-discussed problem of deviant causal chains. Further, I will argue that agents who act for reasons need not consider reasons in a process of deliberation, and they need not treat some consideration as a reason. Finally, I will argue that we do not have reason to believe in our having free will, under the assumption that incompatibilism about free will is true. I will suggest that this result is less drastic than one may think, because there are viable accounts of moral responsibility and of free and autonomous agency available, which do not presuppose libertarian free will.

Note that the three mentioned challenges to the standard-causal model are global or external, rather than internal challenges—that is, unlike the logical connection argument or the challenge from deviant causal chains, they do not point towards internal problems and inconsistencies. Rather, they question the very approach—they

question the idea that agency can be understood in event-causal terms. Accordingly, my responses concern the big picture, as it were, rather than the details of the view. That is reflected in the fact that I will stay neutral on many issues concerning the ontology of action. I will stay neutral, for instance, on the questions whether actions and events are to be individuated finely or coarsely. The general metaphysical framework of this thesis can be described as broadly Aristotelian. I will assume that there are substances—that is, things and beings that persist through change—and properties. I will remain neutral on the question of whether events are particulars or instantiations of properties.

Further, I will defend causalism as non-reductive physicalism, because there is, as I think, *prima facie* reason to endorse that view. But I will not reject other options, such as the identity theory or functionalism. My aim, in other words, is to defend the standard-causal *approach* as broadly as possible, rather than a particular version of it. Apart from the mentioned reasons, my motives for that defence can be summarised as follows.

Firstly, I mentioned two intuitions that count against the standard-causal approach. There are, however, equally strong—if not stronger—intuitions in support of both the standard-causal account of action and the causal theory of reason-explanation. Suppose you recall, while you are reading these lines, that you want to send an important parcel today, and that you realise that the post office will close shortly; and suppose you find yourself on the way to the post office a few minutes later. It is very plausible to think that those thoughts motivated and resulted in your action—that those mental states and events made you to or moved you to head towards to the post office. Given that, it seems quite natural, and only one innocuous step further, to suggest that those mental states and events are among the causes of the action. Similarly, when we ask you why you are heading that way, you will probably say that you are heading to the post office, because you want to send that parcel—and that you have to hurry because the office will close shortly. And given the relevant information, we will explain your action to others in a similar fashion. Again, it seems quite natural to think that such explanations are causal explanations, because it seems that they are true only if the mentioned mental states and events actually caused you to perform the action.

My second motive is integral to my defence. I think that the theory is attractive, because it provides an integrated account of agency in the sense outlined. There are two perspectives on our role and place in the world, which have sometimes been described as being incompatible or in tension with each other.[6] On the one hand, we are active beings that can make a difference to the course of events. On the other hand, we ourselves are part of the event-causal order. Our physical movements, which seem to constitute our actions, have sufficient neural antecedents, which in turn have another sufficient physical cause. A model of agency that can reconcile those two perspectives has a great theoretical advantage over any alternative theory that fails to do so. The standard-causal model, as I will argue, can establish this reconciliation—or integration—, whereas non-causal alternatives fail.

One may interpret the previous point—my second motive—as saying that the standard-causal model is to be preferred, because it confirms to the philosophical orthodoxy—or prejudice—known as *naturalism*. Let me say two things in response to that. Firstly, if my commitment to the standard-causal model is motivated by commitment to naturalism, then the variety of naturalism in question is a very weak one. My motive presupposes only that human agents are part of the order of events in the sense outlined. Naturalism, however, makes typically much stronger claims in addition to that. It says, for instance, that all facts supervene on natural or descriptive facts, that all concrete phenomena must be explained in terms of efficient causation, or that all explanations must be reduced to the natural sciences.[7] Secondly, my motive is not simply that the standard-causal model confirms to some variety of naturalism, and I will not, for instance, reject alternative models for the reason that they assume non-naturalistic kinds of causation, such as final-causation or substance-causation. Rather, my motive is that the standard-causal model is the only model that provides an integrated account of agency by locating action in the event-causal order and by explaining the relationship between, on the one hand, actions and reasons for actions and, on the other hand, physical movements and their physical causes. (I will develop this point in more detail in chapter two.)

---

[6] Compare, for instance, Melden, 1961; Thomas Nagel, 1986; and Bishop, 1989. Or one might as well think of Kant's doctrine of the two standpoints (compare the *Groundwork*, p. 62, for instance).

[7] It is, notoriously, a difficult task to spell out what naturalism says exactly. Compare, for instance, MacDonald, 1992, and Pettit, 1992.

Chapter One: Agents and Their Powers

The main focus in this chapter is on what I shall call reductionism and non-reductionism about agency. At the heart of our conception of agency is the idea that agents are capable of self-movement. What distinguishes agents from other beings or things is that they can bring about change by bringing about change in themselves—they can move things by moving themselves. Reductionism about agency says, very roughly, that an agent's power or ability to engage in agency is reducible to relations between changes in—and states of—the agent; it is reducible, as I shall say, to relations between agent-involving states and events. Non-reductionism about agency denies that. It says that an agent's power—and the relation that holds between the agent and an action—is primitive and irreducible.

In the first part, I will begin with some preliminary remarks concerning the nature of behaviour, actions and agents. I will then say more about reductionism and non-reductionism, and I will introduce a third position—namely, volitionism. Then I will turn to the theory of agent-causation, which is the most prominent version of the non-reductive approach. Proponents of that theory argue that the reductive approach fails to capture important aspects of human agency, that their non-reductive theory can account for those aspects, and that we have therefore good reason to endorse their view. I shall defend the reductive approach against that challenge, and I will argue in this and in subsequent chapters that there is reason to prefer the reductive approach.

In the final section I will, firstly, introduce a fourth position, which I shall call pluralism. Secondly, I will present objections to volitionism. Then I will show that the arguments against the agent-causal theory can be generalised—that they are arguments against the non-reductive view in general. And, finally, I will show that apparently alternative positions fall under reductionism, non-reductionism or pluralism about agency. The apparent alternatives that I will consider are the theory of agency that is implicit in a broadly Kantian moral psychology and so-called emergentist views.

Some Preliminaries on Behaviour and Action

It is a commonplace in the philosophy of action to approach the question what action or agency is by contrasting the things that agents *do* with things that *happen to* them.

What is the difference between, say, Sue's walking down the corridor and her tripping over the doorstep? What is or what constitutes the difference between being *active* and *passive*?[1]

That distinction may be helpful in introducing the subject matter, but it is questionable whether it is useful beyond that. A first problem is that there are many cases, which apparently do not fall under either category. Considers Sam's writing an e-mail, Sam's blushing in response to an embarrassing joke, and Sam's catching the flu. Writing the mail is something that Sam does, and catching the flu is something that happens to him. But what about Sam's blushing? On intuitive grounds, we would neither classify it as something that Sam does, nor as something that happens to him. One may argue, on theoretical grounds, that Sam's blushing can be subsumed under one of the two categories. But such arguments, I think, will not convince us that we *should* subsume the blushing under one category rather than the other. Rather, they will show, at best, that things can coherently be divided up and subsumed in a certain way. Alternatively, and more plausibly perhaps, a third category can be introduced that covers all—or at least the most important—cases that lie between actions and happenings. One might distinguish, for instance, between actions, mere behaviour and happenings, and Sam's blushing could then be classified as an instance of mere behaviour.

However, it is also a commonplace in the theory of action to point out that such classifications are to some extent arbitrary and relative to interest.[2] Any plausible theory of action must have the resources to distinguish between cases like Sam's writing the mail and Sam's blushing. If the latter is classified as an action, then the theory will, presumably, distinguish between at least two very different *kinds* of action; namely, one that subsumes the writing and one that subsumes the blushing. I think nothing of importance depends on whether a theory subsumes it under action, behaviour, mere behaviour, or whatever, as long as the theory has the resources to distinguish it from other kinds of action, such as intentional action and action that is done for reasons.

---

[1] Berent Enç, for instance, distinguishes between two problems of action theory, the first of which is to distinguish 'the class of things that we do as rational agents from things that happen to us' (compare Enç, p. 39-40). See also, for instance, Melden, 1961, especially chapter 6; Davidson, 1980, p. 43; Goldman, 1970, pp.70-71; Frankfurt, 1988, p. 69; Brand, 1984, p. 1; and Dretske, 1988, p. 1.

[2] Compare, for instance, Brand, 1984, p. 4, and Dretske, 1988, pp. 6-7.

In many disciplines the terms *action* and *behaviour* are presupposed as primitive, as are other actional or agential notions such as *response* and *control*. In fact, often the terms *action* and *behaviour* are used interchangeably in one discipline, and differently in different disciplines. But that, of course, is not a problem, for the physicist's talk about the *behaviour* of electrons certainly does not interfere with the philosopher's distinction between behaviour and happenings.

Given that, we should be sceptical about the idea that there is something such as action, agency or behaviour *simpliciter*. We should be sceptical about the idea that we can define those notions in a way that satisfies all our intuitions concerning those notions in all contexts. Rather, we should focus on the features of the kind of agency that we are interested in. Here we are, first and foremost, interested in *human* agency—in particular, the features and aspects of agency, which are distinctive of human agency. It may, of course, be very useful to consider the features of the kinds of agency exercised by non-human animals and organisms and to contrast them with the features of human agency. But, generally, a theory of action must have the resources to distinguish between the relevant features of agency only within its domain of interest. That is to say that it should it not be expected that the theory provides conditions and definitions that satisfy all intuitions concerning the concepts of behaviour and action *simpliciter*.

Adult human agents are—normally or usually—capable of intentional, rational, deliberative, autonomous and free agency. Human agency comprises all those different forms or kinds of agency. In chapter 4 I will say more about the more refined or higher aspects of deliberative, autonomous and free agency, and I will turn to the question of how those kinds or aspects of agency are related. For now, I shall follow a common practice and associate *action* with intentional and rational action in the following sense. Firstly, actions are goal-directed, purposive or motivated in the sense that agents perform them in order to achieve or attain certain goals or ends. And secondly, actions admit of so-called *rationalising* explanations, which are essentially formulated in psychological or intentional terms, and which render the performance

of the action intelligible by referring to some of the agent's mental states and events that constitute—or that can be associated with—the agent's *reasons* for the action.[3]

Let us now return briefly to the distinction between actions and happenings. There is, it seems undeniable, a genuine distinction between things that agents do and things that happens to them. However, it would be a serious mistake to think that, therefore, actions cannot be explained in terms of—or reduced to—happenings.[4] Without further specification a happening is simply an event, and a happening that involves an agent—that is partly constituted by an agent—is an agent-involving event. It would be a mistake to think that every agent-involving event is something that happens *to* the agent. Trivial counterexamples are given by basic bodily functions. Sweating, for instance, is not something that is properly described as something that happens to me.

In fact, it is common to assume that actions can be reduced to events. The easiest way to see why this is a plausible assumption is to consider the description of an action which features a transitive verb with an intransitive counterpart, such as *raise* and *rise*. Whenever Sue raises her arm, the following two events occur: the event of Sue's raising her arm and the event of Sue's arm rising. The former event entails that Sue performed an action, but the latter does not. The reductive view assumes that every description of an action *entails* that an event of the latter act-neutral or non-actional kind occurred, and that the action is either identical with or constituted by that event. In the following, I shall assume that this is correct—I shall assume that actions can be reduced to act-neutral events.[5]

A weakness of the distinction between the things that agents do and the things that happen to them is that it provides merely a negative characterisation of agency. It tells us only what actions are not—namely, things that happen to us. Given that actions are constituted by events, the way forward is to ask *in virtue of what* act-neutral events constitute actions.

---

[3] In the analytical philosophy of action, the two primary sources concerning the close relationship between action, intentional action and acting for reasons are Anscombe, 1957, and Davidson, 1963 (reprinted as essay 1 in Davidson, 1980).

[4] For instance, Moya, 1990, does not distinguish clearly enough between things that happen to agents and mere happenings—that is, events. The same is true of some passages in Melden, 1961, and Nagel, 1986.

[5] I will return to the relation between actions and events in chapter 2, pp. 99.

Self-Movement and Internal Causes

A common strategy is to characterise action or behaviour in terms of its causal history—its history of production. Fred Dretske, for instance, has proposed a characterisation of behaviour as endogenously or internally produced—that is, caused—movement.[6] All movements are, presumably, performed or executed by some being or system, and they are events which *involve* that being or system in the sense that they are partly constituted by it. What distinguishes a mere movement from a movement that constitutes behaviour, according to Dretske, is the fact that the latter is produced by the being or system itself. The movement of a stone, for instance, is produced by external forces; the stone moves, because it is being pushed or pulled by something else. The movement of a cat's paw, however, is produced either by something external to the cat or by the cat itself. Only the latter is behaviour, because only the latter is internally produced movement. That is, I think, a plausible starting point for capturing the intuition that agency has something to do with self-movement—the idea that agents are able to bring about change by bringing about change in themselves.

Dretske thinks that action is a subclass of behaviour; namely, intentional or rational behaviour.[7] What we obtain is a neat categorisation of action as a subclass of behaviour, and of behaviour as a subclass of movements. However, the offered characterisation of behaviour appears to be too wide. At one point Dretske says that behaviour is 'internally produced movement *or change*'.[8] Events such as taking a breath, sweating, digesting, pumping of blood, it seems, are internally produced *changes*. But, intuitively, we would not classify them as behaviour. Dretske, however, is prepared to say that events of that kind are instances of behaviour.[9] He thinks that every system of classification—at least as far as behaviour and action is concerned—will have counterintuitive implications. Since no system will satisfy all our intuitions, we should not dismiss a theory on that ground, provided that it captures

---

[6] Dretske, 1988, chapter 1.
[7] Ibid., pp. 4-5.
[8] Ibid., p. 3, my emphasis.
[9] Dretske does not consider a distinction between the behaviour of the agent as a whole as opposed to the behaviour of its parts. One could exclude things such as the beating of one's heart on the ground that it concerns the behaviour of a part (compare Goldman, 1970, p. 47). But that would not exclude all counterintuitive cases, for events such as inhaling and sweating are agent-involving events that constitute behaviour of the agent as a whole.

the core of the concept of behaviour. Generally, I tend to agree with Dretske. But let us nevertheless see whether we can find an alternative that preserves the proposed characterisation and excludes at least some of the counterintuitive results.

One may, for instance, distinguish bodily movements as a subclass of agent-involving events, and then define behaviour as internally produced bodily movement. Berent Enç, for instance, has suggested a characterisation of the 'basic behaviour repertoire of organisms' as 'macro units of behavioural outputs', which are triggered by 'higher centres of the organism'. An organism's behaviour repertoire is relative not only to species, but also to individuals. Most human animals, for instance, can walk without actively monitoring each single step, let alone the muscle movements involved. Rather, a decision or intention to walk triggers the 'macro unit' that is associated with walking and that controls the individual steps and muscle movements.[10]

Dretske acknowledges that the attribution of agency presupposes that the being or system in question exhibits an internal structure. In particular, it must possess internal mechanisms or sub-systems that are causally responsible for some of its movements. But since Dretske wants his characterisation of behaviour to encompass the movements of the simplest living creatures and organisms—he talks about plant behaviour—, he would probably reject reference to a 'higher-centre' of the organism and a restriction to movements of body parts.

I said that I tend to agree with Dretske's stance concerning the interest relativity of definitions and classifications. In the present case there seems to be no obvious reason to prefer one of the offered proposals, *as far as* the characterisation of *behaviour* is concerned. Things are different, however, if the aim is characterise action as a subclass of behaviour. Because if action is defined as a subclass of behaviour, we must ensure that the characterisation is wide enough as to encompass all actions. The first worry with Dretske's approach was that it is too wide. But when we focus on the characterisation of action as a subclass of behaviour it seems that it is actually too narrow.

Human action encompasses, possibly, both so-called *overt* and *mental* actions. Overt actions essentially involve bodily movements; they are identical with or constituted by agent's moving their bodies. Mental actions do not involve bodily

[10] Enç, 2003, especially pp. 65-67.

movements; rather, they are identical with or constituted by mental occurrences (by the agent's having of thoughts).[11] Plausibly, things like making a decision or solving a puzzle in one's head are mental actions. In any case, we should not rule out the possibility of mental action by beginning with a characterisation of behaviour and action as a subclass of bodily movement.

At that point, one may wonder why we should characterise action as a subclass of behaviour at all. Why can actions not be defined as a subclass of agent-involving events? According to Alfred Mele, actions are like banknotes and sunburn in the sense that their causal history is essential to their identity.[12] A piece of paper is a banknote only if it has been printed by the right institution, and an irritation of the skin is sunburn only if it has been caused by exposure to sun. But the right causal history is not sufficient. Not every piece of paper that has been printed by the right institution in the right way is a banknote; it must be the right kind of paper that is printed in that way. Likewise, exposure to sun can cause different kinds of skin irritation; in order to be sunburn it must be the right kind of skin irritation that is caused in that way.[13] The same, Mele seems to suggest, holds for actions. Having the right causal history is necessary, but not sufficient for an event to be an action: an event, which is an action, is an action only *partly* in virtue of having the right causal history. It is also necessary that it is an event of the right kind.[14]

Is that analogy sound? Is it possible that there are agent-involving events that are not actions despite having the right causal history? If there are such cases, then the range of agent-involving events must be narrowed as to exclude that possibility.[15] That is why one might want to define actions as a subclass of behaviour. Defining actions as a kind of behaviour, one excludes agent-involving events such as breathing, sweating, and the like, provided that those events are not instances of behaviour. And in order to cover the possibility of *mental* action, one may, accordingly, define a

[11] For more on mental action see Mele, 1997. Compare also Bishop, 1989, p. 195, and Enç, 2003, p. 78.
[12] Mele, 2003, 51-52.
[13] Note that the examples are not without problems. Of course, if the causal history of a banknote involves only the printing process, then history alone is not sufficient. An ordinary piece of paper is not a banknote just because it has been printed in the right way. However, if we include the causal history of the paper itself, things are not so obvious.
[14] Mele says that the causal theory does *not* identify actions with '*nonactional* events', but he does not explain how we can distinguish between actional and non-actional events (without circularity). Compare, ibid., p. 52.
[15] Enç, 2003, for instance, thinks that it is necessary to 'narrow the class of event types that will constitute the non-actional neutral events' as to exclude sneezing, blushing, sweating and so on (p. 75).

subclass of agent-involving events as, for instance, the union of the class of behavioural outputs and the class of the relevant kinds of mental events (such as formations of intentions and judgements). But is it correct that the right causal history is not sufficient? Are there counterexamples to the claim that, in the case of action, the right causal history is both necessary and sufficient: are there instances of agent-involving events, which do not constitute actions despite having the causal history of an action?

What distinguishes action, as specified above, is that it admits of true and justified explanation in intentional terms. Actions can be rationalised by reference to some of the agent's mental states and events. On a widespread causal view, which I shall defend, those mental states and events are also among the causes of the actions. The fact that actions admit of rationalising explanation is thereby reflected in their causal history, and they can, therefore, be characterised as being actions in virtue of their causal history. According to that causal theory, the two mentioned features go hand in hand: if an event admits of a rationalising explanation, then it is caused by mental states and events that render it rational, and if an event is caused by such mental states and events, then it can be explained in terms of reasons. (In the following I shall call those mental attitudes the agent's or agent-involving *reason-states*.)[16]

Blushing, it seems, is also caused by agent-involving mental states and events and it admits of psychological explanation. But such explanations are not of the same kind as intentional explanations of actions. Explanations of actions are not just psychological explanations. They are *reason*-explanations. They cite the reasons for which the agent acted. Blushing, however, is typically not done *for* reasons at all. Further, blushing is, typically, not goal-directed and motivated in the way actions are.

Now, blushing is *typically* not motivated and done for reason. But what if we imagine someone who has sufficiently enough control over his reddening of the face such that his blushing can be motivated or even be done for reasons? That may seem far-fetched. A perfectly familiar and plausible borderline example, though, is crying. An actor is supposed to be able to cry on stage *at will*, as it were. In that case, crying is motivated, goal-directed and even performed for reasons. Given that all that holds

[16] Further, the reason-states and events must cause the action in the *right*, *normal* or *non-deviant* way. I will say more on acting for reasons in chapter 2 and 4, and I will turn to causal deviance in chapter 4.

in virtue of the causal history of the event in question, such an instance of crying counts as an action in virtue of having the causal history of an action. But that is just as it should be—for the actor *is* performing an action.

Whether or not there are counterexamples depends, of course, on how the conditions for the causal history of action are formulated in detail. In particular, it depends on whether all the mentioned features—being goal-directed, motivated and done for reasons—can satisfactorily be characterised in causal terms only. It will emerge in subsequent chapters—especially in chapters 2 and 4—that this can be done. For the moment, note only the following two points. Firstly, it is difficult to see how and why examples of blushing, sweating, crying, and so on, could constitute counterexamples. Either they lack certain features of actions, which are features of their causal history, or they possess all those features, in which case it *is* plausible to say that they are actions—such as in the case of the actor's crying on purpose. That suggests, secondly, that in the case of action the right causal history is both necessary and sufficient—what matters, in other words, is not what *type* of agent-involving event is produced, but *how* it is produced. It may still be informative to specify which types of agent-involving events typically constitute actions. But, as far as I can see, it is not necessary to do that.

We started with Dretske's characterisation of behaviour as internally produced movement or change. So far, we have been concerned mainly with what is produced—with behaviour and action construed as the effect or output of that process. In the next section, I will turn to the question what it means that behaviour or action is produced *internally*, which will lead us, eventually, to the two positions that are the focus of this chapter—namely, reductionism and non-reductionism about agency.

Reductionism and Non-Reductionism about Agency

The talk of agent-*internal* causes gives expression to the idea that agency has something to do with *self*-movement—the idea that beings, which are capable of agency, are able to bring about change by bringing about change in themselves. Reformulating the proposal, Dretske says that behaviour has its 'causal origin *within*

the system whose parts are moving'.[17] By that, though, he does not mean that the causes of behaviour and action must literally originate within the agent, in the sense that they are *uncaused* agent-internal causes—or *first* causes within the agent. What does it mean, then, that behaviour is caused *internally* or that the causes are agent-*internal*? We can distinguish between the following three interpretations, which provide us with three main positions in the metaphysics of agency.

According to the first, action or behaviour is produced internally in the sense that it is caused by agent-involving states and events. That is the kind of causal theory aspects of which have already been introduced in the previous section. On that view, there is no need to assume that actions literally originate within the agent, or that they have their causal origin within the self. Rather, it suffices that action or behaviour is caused by appropriate agent-involving states and events in the right way.

According to the second position, action or behaviour is internally produced in the sense that it is caused by an agent-involving mental *action* of the right kind. Most proponents of that view identify—or associate—those mental acts with acts of the *will* and call them, accordingly, *volitions* or *willings*; others talk about *decisions* or *tryings* as the appropriate mental acts. I shall use the term *volitions* in order to refer, generally, to mental acts of that kind. Volitions are, firstly, not caused to occur. They are acts in virtue of intrinsic rather than extrinsic or relational properties.[18] They are *sui generis* acts and a manifestation of spontaneous human agency. And, fourthly, volitions are more basic—or fundamental—than overt actions in the sense that a certain bodily movement is or constitutes an overt action just in case it is caused by a volition.

According to the third position, action or behaviour is internally produced in the sense that it is caused—or performed—by the *agent*, rather than being caused by changes occurring in the agent or states of the agent. The action is caused—or performed—by the being *itself*; by the agent *as a whole*, as it were. According to that view, when an agent brings something about by doing something, the agent, literally, brings it about by bringing about change in itself.

---

[17] Dretske, 1988, p. 2.

[18] Definitions of the notion of *intrinsic* properties usually appeal to duplicates of individuals: a property *P* is an intrinsic property of *s* if and only if *s* has *P*, and for all molecule-to-molecule duplicates *x* of *s*: *x* has *P*. In the present context, the important point is, simply, that causal properties are extrinsic or relational properties.

The second and the third position take seriously the idea that action and behaviour has its *origin* within the agent. Both positions are, usually, held only with respect to *human* agency or with respect to distinctively human aspects or kinds of agency, such as rational, free and autonomous agency. Human agents, it is generally agreed, are capable of rational, free and autonomous agency. According to the second and third position, that requires that actions are determined by the self, either in the sense that the agent *herself* is its origin or in the sense that it is caused by an uncaused agent-involving volition.

The second position, which I shall call *volitionism*, has been widely rejected for the reason that it is subject to a vicious regress. At the end of this chapter I will return to volitionism and the regress objection, and I will explain why I think that the view should be rejected. I shall therefore focus on the first and the third position, which I shall call the *reductive* or *standard*-causal model and the *non-reductive* or *agent*-causal model of agency, respectively. Some clarifications and remarks are in place.

Note, firstly, that all three positions identify a *process*—that is, they identify an instantiation of a relation that holds between an agent-involving event and an antecedent (in virtue of which that event is or constitutes an action). Let us call the former the effect-component, and the latter the antecedent-component of the process. The offered positions suggest that action is identical with or constituted by one of the *components* of the process. Alternatively, one might identify action with the whole process, rather than one of its components. In chapter 2, however, I will argue that there is no reason to endorse the process view, whereas there is some reason to prefer the component view.[19] At this point, note only that the problem of adjudicating between reductionism, non-reductionism and volitionism is independent of the question of whether actions are constituted by components or processes. At least, I fail to see a significant connection between the two issues.[20]

Secondly, note that both the standard-causal and the agent-causal model are compatible with the kind of reductionism presented in the previous section, which reduces actions to events—or event-causal processes. But the standard-causal model

[19] See chapter 2, pp. 96.

[20] Compare Clarke, 2003, for instance, who says that nothing of substance depends on whether one assumes a process or a component view (p. 25 and p. 138). He opts for a component view—or product view, as he calls it—but he says that this choice is 'largely arbitrary', and that the issue is 'little more than a verbal matter' (p. 25).

is a form of reductionism in the further sense that it reduces an agent's role in the production of action to the causal roles of agent-involving states and events. The agent-causal model rejects that kind of reduction, which is why I call the former the reductive and the latter the non-reductive model. Only volitionism is non-reductive in both ways.[21]

Thirdly, note that both the standard-causal and the agent-causal model are, strictly speaking, two *versions* of two positions. The two positions are the reductive and the non-reductive model of agency. The former affirms and the latter denies that the agent's role in the production—or performance—of action is reducible. The two outlined *versions* construe that role as a *causal* role; namely, as causation by agent-involving states and event and as causation by the agent. I call the first the *standard*-causal model, because that approach has been the standard approach in the analytical philosophy of mind and action.[22] And secondly, because it presupposes and incorporates what can plausibly be called the standard view in the metaphysics of causation, according to which all causation is efficient event-causation.[23] The introduced version of the non-reductive view, on the other hand, is called the *agent*-causal model, for the obvious reason that it assumes causal relations between actions and agents.[24] Most proponents of that position identify the agent, which is the agent of *human* agency, with the persisting human animal or organism—with the biological substance that is the human agent or person.[25] Further below I will say more about the

---

[21] Note that this corresponds to the fact that both the standard-causal and the agent-causal model characterise actions in terms of their extrinsic or relational properties—namely, in terms of their causal history—, whereas volitionism typically captures agency in terms of the intrinsic properties of mental actions. I will say more on that at the end of this chapter.

[22] Compare for instance Velleman, 2000, who calls it the 'standard story of human action' (pp. 5-7 and p. 123) and Searle, 2001, chapter 1. Sometimes that view is called the *desire-belief* theory of action. That label is misleading, though, since the theory may refer to mental attitudes other than desires and beliefs as the relevant agent-involving mental attitudes.

[23] That is why the view is sometimes called the *event*-causal theory of action. That is slightly misleading as the causal history of an event might consist of events, states and other standing conditions. To allow states and standing conditions as causes, however, is not to deny the central role of events, which are the entities that trigger or initiate effects.

[24] According to Chisholm, 'the philosophical question is not […] the question whether or not there is "agent causation." The philosophical question should be, rather, the question whether "agent causation" is reducible to "event causation"' (1977, p. 622). The introduced distinction between standard-causal and agent-causal theories is in line with that. The difference is merely terminological, because I use the term *agent-causation* to refer to the position that causation by an agent is irreducible.

[25] I use the term *substance* in the broadly Aristotelian sense as referring to something that *persists* through qualitative change. The view that the agent is the persisting human organism—a biological substance—is, according to O'Connor, the 'official' view (O'Connor, 2002, p. 343). Many proponents of agent-causation think that their position is at least continuous with naturalism. In particular, they

different versions of the non-reductive model. For the moment, note only that the agent-causal theory is *a* version of the non-reductive approach, whereas the standard-causal model can be considered as *the* version of the reductive approach. That is because proponents of the non-reductive model may hold that the relationship between agents and actions is not reducible to event-causation, but deny that the relationship is one of *causation*—they may hold that agents *perform* rather than produce or cause actions. However, if one thinks that the agent's role can be reduced to relations between agent-involving states and events, there is no reason to deny that actions are *caused* by agent-involving states and events.

The Standard-Causal and the Agent-Causal Model

Both the reductive and the non-reductive position are compatible with a causal approach in the theory of action—an approach that construes agency as a causal phenomenon. When agents act, they bring something about—they cause something, when they do something. According to the causal approach, that is not only a claim about the consequences or outcomes of actions, but about the actions themselves. When agents act, they cause their actions. In particular, agents cause their so-called *basic* actions—or the basic act components of their actions. Very often, when agents act they do something *by* doing something else. For instance, one gives a signal by raising one's hand. *Basic* actions are, roughly, the things an agent is able to do without doing something else. They are things that one can do *readily*—and in particular, without using knowledge about how to perform the action *by* doing something else.[26] When Sue raises her arm in order to give a signal, she brings it about that her arm goes up in order to give a signal. Sue's *raising her arm* is the basic act component of the non-basic action of *giving a signal*, because she does not have to *do* anything else in order to raise her arm. It is fairly uncontroversial that there are basic actions and that non-basic actions are, in some sense, *generated* by basic actions.[27] I shall assume that that is correct.[28]

---

think that the theory is not committed to substance *dualism*. Compare Taylor, 1966; Chisholm, 1977; Van Inwagen, 1977; O'Connor, 2000; and Clarke, 2003.

[26] See, for instance, Danto, 1965; Goldman, 1970; Davidson, 1980, essay 3; Ginet, 1990; and Enç 2003.

[27] Goldman, 1970, distinguishes between four different ways in which actions can be generated by more basic actions. Ginet, on the other hand, proposes a 'general generating relation' (1990, pp. 19).

So far, I have talked about agent-involving events in order to refer to both the effect-component and the antecedent-component. It must be noted, though, that some of the things that agents do may not be constituted or generated by *agent-involving* events. Some actions and consequences that are caused or otherwise generated by the performance of basic actions may not involve the agent, in the sense that the events that constitute those actions are not instantiated by the agent—they are not changes in that agent. Suppose that Sam intentionally sets off an alarm by unlocking a door (which presupposes, of course, Sam knows that unlocking the door will set off the alarm). *Setting off the alarm* is something that Sam does and which, arguably, does not involve Sam in the same sense as the movement of his hand involves him.[29] In the following, though, I shall ignore all complications that arise in connection with the distinction between basic and non-basic actions. All statements concerning agent-involving events, as the constituents or antecedents of actions, must be understood as being, *strictly speaking*, about the constituents and antecedents of *basic* actions. That is unproblematic for the following reason. We are interested in the phenomenon of agency—in particular, in human agency and the role of human agents in the performance or production of actions. The distinction between basic- and non-basic actions is uncontroversial. Further, it is generally assumed that non-basic actions can be defined *recursively*, on the basis of a definition of basic actions, by adding clauses concerning the generation of non-basic acts.[30] Hence, to consider the relationship

---

Generally, note that generation can be causal, but need not be. In many cases, such as Sam's voting by raising his arm, conventions play a crucial role in the generation of non-basic actions.

[28] Concerning the individuation of actions and events we can distinguish between three views: coarse-grained or minimising views, fine-grained or maximising view, and mixed views. According to the coarse grained view, actions are particulars (that is, events) with indefinitely many properties (see, for instance, Anscombe, 1963; Davidson, 1980; and Enç 2003). In the example, it is one and the same particular that is a raising of one's hand and a giving of a signal. According to the fine-grained view, two event-tokens are identical only if they involve one and the same property. So, being a raising of one's arm and being a giving of a signal constitute distinct tokens (see, for instance, Goldman, 1970). It is common now to say that the dispute between the two positions is merely verbal, and that it is therefore unproblematic to remain neutral (compare Enç, 2003, note 11, p. 85, and Mele, 1992, pp. 4-5). What is important to note is, firstly, that both views are compatible with both the standard-causal and the agent-causal model. And secondly, proponents of both positions may use the notion of *generation* to formulate their view. Enç, for instance, thinks that non-basic actions are generated or made the case by basic actions, *and* he thinks that they are identical with them (if they are generated in that way).

[29] Whether or not there are actions that do not involve agents at all depends on how actions and events are individuated. But since I remain neutral on that issue, I have to explain why the possibility of such actions is unproblematic.

[30] For recursive definitions of non-basic act-tokens on the basis of definitions of basic act-tokens see, for instance, Goldman, 1970; Ginet, 1990; and Enç, 2003.

between agents and non-basic actions would complicate the issue without giving us any additional insight concerning the agent's role in the performance of action. We can therefore restrict our investigation to basic actions, and I shall continue to talk about agent-involving events as the antecedents and constituents of actions.

I shall defend the reductive standard-causal model of agency against various objections and challenges from alternative approaches. In this chapter I will first argue against the agent-causal theory in particular, and later on I will show that the objections apply to the non-reductive model in general. Before that, though, I shall say more about reductionism.

Reductionism and the Agent

The reductive model reduces not only actions, but also the agent's role in the performance of actions. That raises the following two questions. The view reduces the agent's role to the causal roles of agent-involving states and events. Does that mean that the view literally reduces the *agent* to a sum or system of mental states and events? Is the view committed to the claim that a human agent—or a person—is identical with a sum or system of mental states and events?

Secondly, according to the non-reductive view, reference to agents—and to their irreducible agential powers—is necessary for complete and true explanations of actions (or, in general, for a complete and true description of the world).[31] The reductive view denies that. However, the reductive view talks about *agent*-involving states and event—as being both the antecedents and constituents of action. Is the view threatened by circularity? Is reference to the *agent* of actions necessary, after all?

Note, first of all, that the two questions are closely related in the sense that one may reject what is suggested in the second by answering the first in the affirmative: one may avoid circularity—or necessary reference to the agent—by identifying the agent with a sum or system of mental states and events.[32] Let us call that view reductionism about the *agent*.

---

[31] Compare, for instance, Clarke, 2003, p. 180: 'completely characterising what happens in the world—saying all there is about what brings about what—will require reference to agent causation.'

[32] It must be noted, though, that one can avoid reference to the agent in that way *only if* the individuation of mental states and events does not require reference to the agent. It might be argued that mental states and events are necessarily states and events of an agent—that they must be instantiated by an active thinker or a person. I assume here that this is not so. But that is not to say, firstly, that mental states and events are independent existences. It may well be true that mental states

In her book *The Bounds of Agency*, which is on the problem of personal identity, Carol Rovane argues that, before we ask what the identity or persistence conditions of persons are, we should settle the question of what kind of thing we are dealing with—that is, the first task is to define personhood. Rovane thinks that personhood must be accounted for in normative terms; she argues, in other words, that *person* is a *normative* kind. On her view, where there is a person, there is a rational point of view that satisfies certain standards. The normative standards in question specify what it is to engage in, what Rovane calls, 'agency regarding relations'. Persons are beings whose engagement with others is informed by 'thoroughgoing regard for their rational point of view'.[33] And a rational point of view is, according to Rovane, nothing else than a set of mental episodes that satisfies certain standards concerning unity, long-term planning and future commitments. Whenever a person thinks about—or refers to—herself, the first person pronoun refers to a 'set of rationally related intentional episodes via a token-reflexive thought that belongs to the set.'[34]

What is interesting for us is, firstly, that according to Rovane a person is, metaphysically speaking, nothing but a *set* of mental episodes, and secondly that all persons are agents. Rovane's theory presupposes, naturally, that persons are capable of engagement and interaction with others; they are rational and reflective agents with certain capacities and attitudes.[35] All persons are agents, but, possibly, only some agents are persons. Since persons are nothing but sets of mental episodes, it follows that *some* agents are nothing but sets of mental episodes. Namely, those agents that are persons: every agent that is a person is, metaphysically speaking, nothing but a set of mental episodes.[36]

---

and events must be instantiated or owned by some being—say by an organism or biological substance. What is denied is only that they must be instantiated or owned by a being that is necessarily a human agent—or a person. And secondly, that does not mean that mental states and events can exist in isolation, as it were. That is, it may well be true that mental states and events are necessarily parts of a rather complex system of mental and states and events—the view, in other words, is compatible with *holism* about mental attitudes.

[33] Rovane, 1998, p. 88.

[34] Ibid., p. 223.

[35] Ibid., chapter 3.

[36] Rovane maintains that Derek Parfit defended that kind of reductionism in his *Reasons and Persons* (1984)—a kind of reductionism according to which 'persons are *nothing but* certain sorts of episodes standing in certain sorts of relations' (Rovane, 1998, p. 134).

Rovane's account of personhood is committed to reductionism about the agent. *Prima facie*, that position is unattractive and implausible.[37] I shall, however, not argue against it. Instead, I will show that the reductive model of *agency* is not committed to reductionism about the *agent* by providing an alternative proposal that avoids necessary reference to the agent, and that does justice to our intuitions, as it identifies the agent with the being that performs the bodily movements, which constitute its overt actions—on that view the agent is, simply, the being that moves around.

The challenge is to avoid reference to the agent without literally reducing the agent. The alternative proposal says, basically, that there is no problem of circularity and that the challenge is misguided, because the talk about agent-involving states and event is *elliptical*. Both the effect-components and antecedent-components of the identified processes are agent-involving states and events. As explained above, the effect-components—those events that are actions—are either behavioural outputs or mental events of a certain kind. The individual being or system, which has or instantiates such agent-involving events, is, therefore, a being or system, which is capable of producing or performing behavioural outputs and of forming mental attitudes, such as intentions and judgements. But, according to the reductive model, that individual is not yet an *agent*. It is an agent only if those agent-involving events are sometimes caused—in the right way—by reason-states. That means that by referring to *agent*-involving events, we are, strictly speaking, *not* referring to events that involve an *agent*. Rather, we are referring to an individual that has or instantiates those events—events that are actions only if they are caused and rationalised by mental states and events, which are owned and instantiated by the same individual.

[37] It might be instructive to compare reductionism about the agent with what is known as the Humean bundle theory of the self (Hume, 1888, especially p. 252). It says, very roughly, that mental events are not owned by a self or subject, but that they are part of a bundle or collection of mental events, and that the self, therefore, is nothing but a bundle of perceptions or experiences—rather than a substance (compare Shoemaker, 1963). There are, obviously, striking similarities between the bundle view and reductionism about the agent. One point of difference is that the latter talks about the agent or the *active* self, whereas the Humean view is apparently about a perceiving or *passive* self. That is why the Humean bundle theory may appear more attractive than reductionism about the agent. However, we do not want to say, presumably, that the active and the passive self are distinct entities. More plausibly, activity and passivity are two aspects or modes of one and the same self. So, if one reduces the self to a bundle, sum or system of mental states and event, one reduces thereby both the passive and the active self. Compare, for instance, Owen Flanagan's suggestion that 'self-representation involves thinking […] in either of two modes: in the active agent mode or in the passive object mode—as seer or as seen' (1992, p. 178).

Something very similar holds for the agent-involving mental states and events, which are the causal antecedents of action. The agent-involving states and events in question are mental states and events, which rationalise the action and which are owned or instantiated by the same individual that instantiates the action. Again, to refer to that individual as the owner of those mental states and events is not necessarily to refer to an agent. That individual is an agent only if those mental states and events cause and rationalise actions. So, again, by referring to those agent-involving states and events, we are, strictly speaking, *not* referring to states and events that involve an agent. But we are referring to the individual that owns or instantiates those states and events in question.

Given that, reference to agent-involving states and events is not necessarily reference to an agent. Rather, it is reference to an individual that owns or instantiates those states and event and that *is an agent in virtue of* having mental states and events, which sometimes cause and rationalise events that are instantiated by the same individual.[38] (That proposal can be expressed a more straightforward manner, if we assume, for the moment, that actions are a subclass of behaviour, and that we can characterise behaviour independently. We can then say that agent-involving states and events do not literally involve an agent, but an individual—being or system—that is capable of behaviour. Accordingly, that individual is or counts as an agent in virtue of having mental states and events that sometimes cause and rationalise its behaviour.)

The reductive standard-causal model of agency is reductive not only in the sense that it reduces actions to events, but also in the sense that it reduces the agent's power or role in the production of action. The proposal just presented shows that it can do so without circularity; that is, without presupposing reference to *agents*. In the following,

[38] There are, it seems, two problematic kinds of cases. In the first, the agent brings about change in its environment by means of a *prosthetic device*, which is, say, connected directly to the relevant nerve endings (as opposed to being controlled by, say, a remote control, which is controlled by the agent). The second kind of case is *collective* agency, where distinct individuals collectively decide and act in a way such that the collective can be held responsible. Concerning the device, I think we can allow a spectrum of cases that ranges from cases in which the device is a mere *tool* to cases in which the device is a *part* of the agent (such that the agent is the mereological sum of the device and the being that instantiates the antecedents of actions). Whether the device is a tool or a part of the agent depends on the details. (Is the connection between the mental antecedents and the movements of the device reliable? Does the agent take responsibility for the actions that are brought about by the movements of the device?) With respect to collective agency, I am not aware of a convincing argument to the conclusion that collective agency is on a par with—or more fundamental than—individual agency. Given that, it seems very plausible to regard collective agency as *derivative*—as a kind of agency that presupposes agency exercised by individuals.

though, I shall continue to talk about agent-involving states and events as the causes of actions. One must keep in mind that this is elliptical in the sense explained. Those states and events do not literally involve an agent, and they cause actions in the sense that they cause events, which are actions in virtue of having the right causal history.

Note that according to that proposal the agent *is* the individual, but the individual is not essentially an agent.[39] It is an agent during some period—or periods—of its existence in virtue of having certain accidental properties. If it loses some of those properties, which are constitutive of its being an agent, it ceases to be an agent, but it may well be the same individual being.[40]

Reductionism and the Self

Agency has something to do with self-movement. Generally and fundamentally, that may be taken to mean that agents—beings capable of agency—are capable of bringing about changes in their environment by bringing about changes in themselves (either as a response to changes in their environment or in pursuit of some of their goals). With respect to human agency, however, one may think about the related notion of self-determination—self-governance or autonomy—in connection with that. What is at the centre of the concept of human agency is that it is sometimes *up to oneself* to choose what to do—freely and in the light of reasons. Human agents can sometimes determine what to do, without being determined to do so—sometimes the *source* or *origin* of their actions lies within themselves.

The non-reductive agent-causal model provides a straightforward interpretation of the notion of self-determination as a causal relation that holds between the action and the self (the human agent or person). An agent exercises the ability of self-

[39] Note that the individual being is not only the subject of mental states and events, but also a being capable of behavioural output. In the case of a human agent, the agent is therefore not to be identified with the human brain.

[40] It seems clear that non-reductionism about agency must reject reductionism about agents. If the agent's power or role in the performance of action is not reducible, the agent cannot possibly be a sum or system of mental states and events. It is not clear, however, whether the non-reductive model is compatible with the view that some individuals are agents in virtue of some of their accidental properties, or whether it is committed to the much stronger view that those individuals are *essentially* agents. Timothy O'Connor, a proponent of the agent-causal theory, thinks that an adequate account of human agency requires that human agents are 'among the truly basic entities whose activities determine the way the world is' (O'Connor, 2000b, p. 115). But what does it mean to be among the truly basic entities? One might think it means that *we* are *essentially* human agents. Construed in that way, non-reductionism about agency seems incompatible with the view that some individuals are human agents in virtue of their accidental properties.

determination with respect to some action if and only if the agent agent-causes that action. On that view, self-determination is literally causation by oneself as an agent.

But it is by no means obvious that the notion of self-determination must be understood in that way. In particular, it is not obvious that it must be understood as a relation between an action and the self at all. In the last chapter I will outline an alternative account of autonomous agency that is compatible with the reductive model of agency.[41] For now, consider only the following remark concerning the notion of the self.

Instead of asking what kind of thing the self is, or whether there is such a thing as the self, one can ask in virtue of what features an agent is or counts as being or having a self. Obvious starting points are self-referential mental abilities and properties. An agent can be a self in the sense of being aware of itself as an agent, and in the sense of having the capacity to refer to itself and to reflect on its own agency. Further, an agent can have a self in the sense of having a conception of itself—a conception of what kind of being it is, where it comes from, what it is up to, and so forth. According to an account of that kind, the self is the same individual being as the agent. It is one and the same individual that is an agent in virtue of having certain agential properties and abilities, and that is a self in virtue of having certain self-referential properties and abilities. But what is interesting, as far as self-determination is concerned, is not the fact that the self is the same being as the agent, but in virtue of *what* properties and abilities that being counts as being or having a self. What is interesting, in other words, is not the fact that the autonomous agent *is* the agent, but in virtue of what properties it is an *autonomous* agent and what role those properties play in the performance or production of its actions.

Agent-Causation

According to some proponents of the agent-causal theory, all human actions are caused by agents. According to others, only some distinctively human actions—such as free or autonomous actions—are caused by human agents or persons.[42] The human

[41] See chapter 4, pp. 196.

[42] One can distinguish between more positions concerning the question what is caused by the agent and with respect to what kinds of action the agent-causal power is exercised. I shall, however, distinguish only between two main positions, which is sufficient for present purposes. For discussions of different versions of the agent-causal theory see O'Connor, 2000 and Clarke, 2003.

agent is usually identified with the human animal or organism—with a biological substance.

The theory has been criticised and rejected mainly on the ground that it presupposes an untenable—or at least a very contentious—conception of causation by substances. I shall set aside the question whether the notion of substance-causation is coherent or possible. Rather, I will focus on the motives behind the agent-causal approach. Proponents of the agent-causal model argue that there is reason to endorse their theory—and, hence, reason to assume substance-causation—because no other theory of agency can capture or account for important aspects of human agency. In the following, I will restrict my considerations to the standard-causal and the agent-causal model, firstly because I will reject other alternatives in due course, and secondly because my aim is to defend the reductive standard-causal approach against the agent-causal challenge. Given, then, that the standard-causal and the agent-causal models are the main contenders, we have to distinguish between two lines of argument.

According to some agent-causalists, the standard-causal model of agency is unsatisfactory. They acknowledge that the standard-causal model can capture some aspects of agency. But they argue that it is unsatisfactory in the sense that it fails to account for some crucial aspects of human agency, such as acting with free will. That is the line of argument that I will reject in this chapter.

According to others, the standard-causal model of agency is altogether inadequate. They think that the reductive approach of the standard-causal theory does not capture the phenomenon of agency at all—let alone distinctively human kinds of agency, such as rational and autonomous action. I will say more on that in the next section, but my main response will be given in chapter 4.[43]

Further, we have to distinguish between two versions of the agent-causal theory. The proponents of the model agree that some actions are caused directly by the agent. But there is disagreement over the question whether an agent's reason-states play any role in the causation of action. (Recall that reason-states are those agent-involving mental states and events in the light of which the performance of an action appears as intelligible). According to one version, the agent is the *sole* cause of action—or free and autonomous action. In particular, the agent's reason-states are not among the

[43] See chapter 4, *The Challenge of Disappearing Agency*, pp. 155.

causes of action. According to an alternative version, both the agent *and* agent-involving mental states and events are among the causes of action. In particular, the agent's reason-states can contribute causally to the occurrence of an event that is or constitutes an action. I shall follow Randolph Clarke in calling the former view the *traditional*, and the latter the *integrated* agent-causal view.[44]

Proponents of an integrated model will, presumably, restrict their arguments to the claim that the standard-causal model is unsatisfactory, since they acknowledge that agent-involving states and events play a role in the causation of action—which is a central element of the standard-causal account of acting for reasons. Proponents of the traditional version, however, typically reject the standard-causal approach altogether.

The Traditional Model

Proponents of the traditional model think that agents are the sole causes of actions.[45] They deny that agent-involving mental states and events play any role in the causation of action. The claim that reason-states are causally relevant in the production of action, however, is central to the standard-causal account of rational action—of what it is to act for reasons. Because of that, traditional agent-causalists reject the standard-causal model not only because it fails to account for free will, but also because it misconstrues our ability to act for reasons. They think that whenever we act for reasons, we *freely* choose what to do on the basis of—or in the light of—reasons, and that this feature is incompatible with the assumption that reasons for actions—or reason-states—are among the *causes* of action. Claims of that sort are usually supported by phenomenological reflections on deliberation and the act of choosing for reasons. When we deliberate, we are, usually, aware of various reasons and motives. But, as for instance O'Connor observes,

> within the framework of possibilities [...] that these present conative and cognitive factors set, it seems for all the world to be *up to me* to decide which particular action I will undertake. The decision I make is no mere vector sum of internal and external forces acting upon me during the process of deliberation

[44] See Clarke, 2003, pp. 135.
[45] Proponents within analytical philosophy are, for instance, Chisholm, 1964 and 1975; Taylor, 1966 and O'Connor, 2000.

> [...]. Rather, *I* bring it about – directly, you might say – in response to the various considerations [...].[46]

What O'Connor has in mind, when he talks about the internal 'forces' that 'act upon' the agent, are the mental states and events, which the standard-causal model construes as the causes of rational action. Given that how the agent is going to decide is determined only by mental causes and influences from the environment, rather than the agent, the action appears as a *mere vector sum* of causal antecedent factors. The agent does not have the power to decide *on the basis of* reasons, since the causal power of the agent is nothing more than the balance of the causal forces of the reason-states that 'act upon' the agent. If we take seriously the idea that the agent acts *in the light of* reasons, we have to look for some causal factor other than the agent's reason-states. If the agent has the power to act *on* some of her reasons, she must have a power that is distinct from any combination of the causal powers associated with her reason-states.[47] And only an agent-causal theory, as the argument goes, can account for such a power.

There are three important things to note with respect to that argument. Firstly, if the outlined case from acting for reasons were successful, it would support the stronger claim that the standard-causal model fails to account for agency altogether. According to the standard-causal model, certain events are actions in virtue of being rationalised and caused in the right way. That is to say that all action is rational in the sense that it can be rationalised in the light of some of the agent's mental states and events. In that sense, the model construes action as acting for reasons. If it fails to account for acting for reasons, it fails, thereby, to account for action as such.

Secondly, the argument says, basically, that acting for reasons is necessarily acting with free will: whenever we act for reasons we choose with free will in the light of reasons. I will argue that the agent-causal theory fails to account for free will—just as the standard-causal model does. That argument will also undermine the outlined case from acting for reasons. If it is assumed that acting for reason is necessarily acting with free will, and if the agent-causal model fails to account for free will, then it also fails to account for acting in the light of reasons. (In the

---

[46] O'Connor, 1995, p. 257.

[47] In chapter 4 I will argue that acting for reasons is not necessarily based on deliberation (see pp. 184). There it will become clear that O'Connor's comparison between the standard-causal account of deciding and acting for reasons and 'mere vector sums' is unjustified and inappropriate.

following I shall call the view that acting for reasons is necessarily acting with free will *libertarianism* about acting for reasons.)

Thirdly, proponents of the traditional version deny that reason-states play any causal role in the causation of action. But, of course, they do not deny that reasons influence the actions of rational agents. Reasons do influence our actions, but not casually and not directly. Rather, they influence the actions of rational agents by influencing agents *rationally*.

The standard-causal model provides not only a causal theory of action and acting for reason, but it provides also a causal theory of reason-*explanation*. One problem for the traditional view is to provide an alternative account of reason-explanation. The central claim of the causal theory is that reason-explanation is a species of causal explanation: the reasons, for which an action is performed, cause and causally explain that action.[48] The only alternative to the standard-causal account is a *non-causal*—or *teleological*—theory of reason-explanation. In chapter 2, however, I will discuss and reject non-causal alternatives.[49]

In the following, I will therefore focus on Clarke's integrated agent-causal ftheory, which incorporates the causal theory of reason-explanation. However, the main thrust of my objection applies, as will become clear, to the traditional version as well. So, even though the focus will be on the integrated version, my objection will address the agent-causal approach in general.

Clarke's Integrated Model of Agent-Causation

Randolph Clarke has proposed an integrated model that combines a standard-causal model of action and reason-explanation with the agent-causal theory. The integrated standard-causal model allows for genuinely indeterministic—or probabilistic—causal relations between the agent's reasons and the action, and Clarke acknowledges that it provides a viable account of action and reason-explanation. But that model, as Clarke argues, cannot account for free will and moral responsibility; in particular, it cannot account for the kind of *control* that is required both for acting with free will and for

[48] The most prominent statement of that view can be found in Davidson's 'Actions, Reasons, and Causes', reprinted in Davidson, 1980. For a desire-belief version of the view see, for instance, Davidson, 1980, and Goldman, 1970. For versions in which intentions play an additional and central role see, for instance, Brand 1984; Bratman, 1987; Mele, 1992; and Enç, 2003.

[49] Compare also Clarke, 2003, p. 21-24, where he presents arguments against non-causal accounts of reason-explanation in support of his integrated version of agent-causation.

being truly morally responsible. He thinks that the agent-causal model can account for those features of human agency, and he suggests combining the two models. The integrated view says that whenever an agent acts with free will, and whenever the agent is morally responsible, the action is caused by the agent's reason-states *and* by the agent, such that the reason-states alone render the performance of the action only *probable*.

Let us consider first the standard-causal component. Suppose that some agent, call him Sam, has to decide whether to *A* or to *B*. Let *R* be the set of all reasons for and against *A*-ing and *B*-ing, which Sam considers in his deliberation. Assume that *A*-ing and *B*-ing are genuine alternatives for Sam in the sense that Sam is not coerced, forced or otherwise constrained in making his choice, and in the sense that it is not causally determined whether Sam will choose to *A* or *B*. In particular, assume that *R* leaves it undetermined how Sam will decide and act.

It was a commonplace in the debate on free will to say that indeterminism undermines an agent's control over her choices and actions. The thought was, very roughly, that what is not—or cannot be—determined, cannot be under anyone's control.[50] Clarke, however, argues convincingly that indeterminacy does not undermine control.

Control, as I shall assume with Clarke, is a causal phenomenon. According to a causal theory of acting for reasons and reason-explanation, the explanatory and motivational force of reasons is grounded in a causal connection between reason-states and actions. In the second chapter I will discuss that point at great length.[51] For the moment note only that there is no obvious reason to think that the causal connection between reason-states and action must be deterministic. All that is required is that there is a *causal* connection—it may be deterministic or probabilistic. Given that, there is no reason to deny that in the example Sam chooses and acts *for* reasons in either case, since either alternative will be caused by Sam's reasons for *A*-ing or *B*-ing, respectively. But when an agent acts for reasons, then, clearly, that agent exercises some kind of control (which can be called *agential* control—the kind of

[50] Kane, 2002, summarises that argument as follows: 'An event that is undetermined might occur or not occur, given the entire past. Thus, whether or not it actually occurs, given its past, would seem to be a matter of chance. But chance events are not under the control of anything, hence not under the control of the agent. How then could be free and responsible actions?' (pp. 22-23). Van Inwagen, 1983, discusses and rejects three versions of that argument (pp. 126-152).
[51] See chapter 2, *The Case for Causalism*, pp. 73.

control exercised by agents when they perform actions). That shows that the causal processes that constitute control need not be deterministic and that indeterminacy does not by itself undermine control.[52] Clarke argues, furthermore, that indeterminacy does not even *diminish* control. Whether control is exercised or not is a matter of what *actually* causes what. In the example, whether Sam exercises agential control is a question of whether the action that Sam actually performs is caused by his reason-states, rather than a question of whether his reason-states might have caused the other action instead.[53]

Further, Sam has open or alternative possibilities in an objective or metaphysical sense due to the fact that Sam's reason-states leave it undetermined whether he chooses to *A* or to *B*. However, Sam's agential control is restricted to the kind of control associated with acting for reasons. If Sam decides to *A*, then his action is guided by his reasons for *A*-ing. That kind of control, as Clarke argues, is compatible with determinism. It is therefore compatible with Sam's lacking alternative possibilities in the indicated metaphysical sense. To assume indeterminism gives Sam alternative possibilities, but it does not add anything in terms of control. Sam will exercise control in the sense that both *A*-ing and *B*-ing will be done for reasons, but the indeterministic standard-causal model, as Clarke says,

> fails to secure for the agent the exercise of any further positive power to causally influence which of the alternative courses of events that are open will become actual.[54]

It is necessary but not sufficient to require openness. Rather, it must be shown that the agent also has agential control over *which* of the alternatives will become actual. That is why the *integrated* agent-causal theory assumes in addition to the kind of control that is constituted by the agent's acting for reasons,

> a further power to causally influence which of the open alternatives will be made actual. In exercising this further power, the agent is literally an originator of her action, and neither the action nor her initiating the action is causally determined by events.[55]

---

[52] It might be helpful to note that indeterminacy does not entail randomness, arbitrariness or fortuitousness. Assume that the occurrence of an event e_1 in circumstances C renders the occurrence of e_2 probable. It would obviously be a mistake to say that an occurrence of e_2 is *random* or *arbitrary*, given that e_1 caused e_2 in C.

[53] Compare Clare, 2003, chapter 5.

[54] Ibid., p. 133.

[55] Clarke, 2003, p.151.

The strength of this model—as opposed to the traditional version—is that it integrates the standard-causal account of acting for reasons and reason-explanation. Clarke acknowledges, thereby, that the standard-causal approach can capture an important aspect or kind of human agency. What we have to ask is, firstly, whether the standard-causal theory in fact fails to account for free will and moral responsibility, and secondly, whether the agent-causal model can do any better.

The Case for Agent-Causation from Free Will

An agent who has free will, I take it, has the ability or power to do otherwise. It is common to distinguish the following two necessary conditions.[56]

> (AP) *Open or Alternative Possibilities*. It is open to the agent to decide and do otherwise than she actually does. In the circumstances, the agent could have decided and done otherwise.[57]
>
> (SDO) *Self-Determination as Origination*. The agent *herself* determines the decision and action. The agent is not only the cause, but the *source* or *origin* of her action.[58]

The case for the agent-causal theory from free will goes along the following lines.

(1) Free will is worth wanting.

(2) Causation by a substance is possible.

(3) Incompatibilism about free will is true; that is, both AP and SDO are incompatible with the thesis of causal determinism.[59]

---

[56] Compare, for instance, Kane, 1996 and Clarke 2003. Clarke maintains that the characterisation in terms of alternatives and self-determination is shared by most incompatibilists and compatibilists (for references in support of that claim see note 2, pp. 3-4). I assume that it is possible that the conditions for free will are different from the ones for moral responsibility. Generally, free will is an agential power or ability. According to a traditional characterisation, free will is whatever kind of agency that is necessary and sufficient for moral responsibility: the conditions for free will are the conditions for moral responsibility. According to the approach that I endorse, however, free will can be characterised independently as the power to do otherwise.

[57] This characterisation is supposed to capture an intuition concerning what having free will consists in, and it is supposed to be compatible with all the traditional positions on free will. In particular, the mentioned openness and the requirement that the agent could have done otherwise are not supposed to rule out causal determinism (or any other forms of determinism).

[58] Kane, 1996, offers four conditions for 'sole authorship' or 'underived origination', which 'many ordinary persons believe in': '(i) The source or ground (*arche*) of action would be in the agent or self, and not outside the agent. This would mean that (ii) if we were to trace back the causal or explanatory chains of action backward to their sources, they would terminate in actions that can only and finally be explained in terms of the agent's voluntarily or willingly performing them […]. (iii) The agent would be […] responsible to some degree for the self which was formed by them and for subsequent actions issuing from that self. (iv) These self-forming actions would not be determined by anything within or outside the self for which the agent was in no way responsible' (p. 79).

[59] Kane, 1998, distinguishes between different kinds of determinism, such as physical, psychological, theological and logical determinism. He maintains that the core of the notion of determinism is the

(4) A theory of agency is either a causal or a non-causal theory.

(5) Agency is a causal phenomenon; non-causal theories of agency are inadequate.

(6) There are two viable causal theories of agency: the standard-causal and the agent-causal theory.

(7) Standard-causal theories are inadequate (with respect to either AP or SDO).

(8) Agent-causal theories can satisfy both AP and SDO.

(9) Hence, we have good reason to believe in agent-causation.

Many assumptions in that argument are controversial. I shall assume (1) to (7) for the sake of the argument, as I will argue only against (8). But I shall add some remarks with respect to some of the assumptions before turning to (8).

In chapter 4 I will argue that we do not have reason to believe in our having free will. One part of my argument will show that our intuitions concerning the existence and value of free will are not conclusive. The important point is that I shall *argue* and present *reasons* for those claims. That is, nothing of what I shall say chapter 4 is incompatible with the claim that free will is, *prima facie*, worth wanting.

As already indicated, the assumption that substance-causation is possible—coherent or intelligible—is very controversial. Most objections to the agent-causal model are, in effect, objections to that notion of causation. Many of those objections focus on the fact that substances, unlike events, are not *structured* entities.[60] In particular, they are not *in time* like events; they do not *occur*. That fact makes it rather difficult to understand, firstly, how reference to a substance can be of explanatory relevance. For instance, it is central to the explanatory force of an event-causal explanation that the occurrence of the cause explains the occurrence of the effect—a feature which is lost when the cause is a substance.[61] Secondly, the fact that

---

following: an event is determined just in case there are conditions 'whose joint occurrence is (logically) sufficient for the occurrence of the event' (p. 8). Usually and generally, though, philosophers are concerned with *causal* determinism, which is often characterised in terms of laws of nature and possible worlds: the (possibly complex) causal antecedent *c* causally determines the occurrence of an event *e* just in case in every possible world in which *c* and the actual laws of nature obtain, *c* causes the occurrence of *e*. One may think that incompatibilism requires only that *either* AP *or* SDO is incompatible with causal determinism. That is, strictly speaking, correct. However, incompatibilists typically argue that both conditions are incompatible with causal determinism. Compare, for instance, Kane, 1998, especially p. 75.

[60] Events are structured in the sense that they have objects, properties and times as their constituents. Substances are not so structured. For more on that compare Clarke, 2003, especially p. 204.

[61] That point is independent from the following. The two main theories of causation are broadly Humean and counterfactual theories; the former require law-like connections and latter are about

substances are not structured entities makes it difficult to understand the metaphysics of substance-causation, since causation seems to be essentially *dynamic*.[62] However, I shall not develop or discuss those points further and restrict myself to the observation that agent-causation and substance-causation are problematic notions, since I will argue against the agent-causal model on different grounds.

Assumption (3) says that incompatibilism about free will is true. In contrast to the previous assumptions, I assume that not merely for the sake of the argument. I think that free will, construed as the ability to do otherwise, is not compatible with causal determinism, and I will explain why in chapter 4.[63] For the moment note only that incompatibilism does not assert that we have free will; it says only that having free will is incompatible with causal determinism. If we add the claim that we have free will to incompatibilism, we obtain libertarianism. Hence, libertarian free will is incompatibilist free will, and given incompatibilism, having free will is having libertarian free will.

Assumption (7) says that standard-causal theories are inadequate with respect to either AP or SDO. Depending on whether the causal relations between reason-states and actions are assumed to be probabilistic or deterministic, a standard-causal model will violate either the second or both conditions for free will.

Consider first *deterministic* causal relations between reason-states and actions. This version is directly ruled out by (3). How the agent is going to act is, in the circumstances, causally determined by the balance of the agent's reasons. In the circumstances, which include the agent's reason-states and character traits, there is no objective chance that the agent will do otherwise than she actually does. That violates

---

counterfactual dependencies between *types*. Both theories, it seems, are incompatible with the agent-causal model, which denies any kind of 'constant conjunction' between agents and actions. Humean and counterfactual theories are reductive and anti-realist accounts of causation. They identify causation with connections of a different kind (law-like regularity and counterfactual dependence). Provided, though, that there are viable non-reductive or realist accounts of causation available, the fact that reductive theories are incompatible with substance-causation is not decisive against substance-causation. Compare O'Connor, 1995 and Clarke, 1993.

[62] Proponents of agent-causation may reply that such objections are begging the question. For what those objections say is, simply, that substances cannot be causes, because they are not events—or because they are not sufficiently *like* events. However, the problem is not simply that substances are not like events. One can refer to the differences between substances and events in order to explain why we can understand event-causation, but not substance-causation. But it is the metaphysical nature of substances, rather than those differences, which is at the root of the problem. For surveys and discussion of arguments for and against substance- and agent-causation see, for instance, O'Connor, 2000, and Clarke, 2003.

[63] See chapter 4, pp. 206.

AP—given that incompatibilism is true. It is important to note, though, that the deterministic version can account for self-determination in the sense that the actions are determined by the agent's *own* reason-states. But since the agent-involving desires, beliefs and intentions are themselves caused by circumstances that are external to the self, the action does not originate within the self—the agent is not the origin or source of the action. We get self-determination, but not origination—SDO is violated as well.

Now assume that the causal relations are *probabilistic*, and consider again Sam, who has to decide whether to *A* or *B*. Let *R* be the set of reasons including all reasons for and against *A*-ing and *B*-ing that Sam considers and suppose that the probability that *R* causes Sam's *A*-ing is the same as the probability that R causes Sam's *B*-ing. No matter which action Sam chooses to do, his action will be caused by his reason-states and there is an objective chance that Sam does otherwise—for reasons. Hence, AP is satisfied. But with respect to SDO, indeterminism cannot help. No matter how the agent decides, the decision will be caused by the agent's own reason-states. But, for the same reasons as above, it is not the case that the agent is the origin or source of the action. (Note that there is a close connection between that point and Clarke's point concerning control presented above, pp. 31. The agent is in control in the sense that the action is caused by the agent's reason-states in the right way. But since the reason-states leave the choice undetermined, the agent does not determine—does not have control over—which of the open alternatives will become actual. I shall say more on the problem of control further below, and will return to that issue in chapter 4.)

We can conclude that neither the deterministic nor the probabilistic version of the standard-causal model can account for origination; hence (7) holds. Proponents of agent-causation think that their view can account for both AP and SDO—and for the associated kind of control. For, if the agent has the power to cause an action directly, the agent is, literally, the *origin* or *source* of that action. Given that, they can conclude, in conjunction with the assumption that non-causal theories are inadequate, that there is reason to endorse the agent-causal theory, as it is the *only* theory that can account for free will.

The Case for Agent-Causation from Moral Responsibility

The case for agent-causation would be considerably stronger, if it could be shown that the agent-causal model is the only model that can account for moral responsibility.[64] The presented argument from free will would show that under the additional assumption that free will is a necessary condition for moral responsibility—that AP and SDO are necessary conditions for being morally responsible. (In the following I shall call the view that being morally responsible presupposes libertarian free will *libertarianism* about moral responsibility).

The conditions for moral responsibility, however, are more controversial than the ones for free will. If moral responsibility is compatible with determinism, both deterministic and probabilistic versions of the standard-causal theory might be adequate—as far as moral responsibility is concerned. If it is incompatible, we need an account of agency that satisfies AP. A probabilistic version of the standard-causal model can satisfy this condition. The question would then be whether self-determination *without* origination is sufficient for moral responsibility. If origination is not necessary, the standard-causal theory remains an option.

The case for agent-causation stands here on rather weak grounds. Proponents of the agent-causal theory could easily construct a case from moral responsibility by reformulating the presented case from free will, if both AP and SDO were necessary conditions for moral responsibility. However, Harry Frankfurt argued, convincingly as I think, that AP is not necessary for moral responsibility.[65] Further, since probabilistic standard-causal theories satisfy AP and since they can account for self-determination, proponents of agent-causation must assume that *origination* is crucial to accountability. Further below, however, I will argue that origination does not make a relevant difference, as far as moral responsibility is concerned.

---

[64] Of course, neither the agent-causal nor the standard-causal model provides an account of moral responsibility. What I am assuming here is that an agent's being morally responsible presupposes that the agent exercises a certain kind of agency. The question is, then, whether that kind of agency can be accounted for by the theory of agency in question. We can then say that a theory of agency can account for moral responsibility, in the sense that it can account for the kind of agency necessary for moral responsibility.

[65] For the well-known counterexamples to the claim that alternative possibilities are necessary for moral responsibility see Frankfurt, 1969, reprinted in Frankfurt 1988. For more on that, and for a defence of a compatibilist theory of moral responsibility see for instance Fischer and Ravizza, 1998 and Haji, 1998. Compare also chapter 4, p. 197, note 78.

In his book *Libertarian Accounts of Free Will*, Clarke has questioned the possibility of causation by a substance. But he argues that an integrated agent-causal theory *would* provide the best libertarian account of free will and moral responsibility, *if* substance-causation were possible. Given that, libertarians about free will and moral responsibility have good reasons to endorse the agent-causal approach, and to develop and defend a theory of substance-causation. Against that, I will argue that agent-causation does not help with libertarian free will and moral responsibility. (Note that from that it follows that the agent-causal theory does not help with libertarianism about acting for reasons either.)

Explanation and Control

Reference to an event-cause can explain why a certain type of effect occurred when it occurred. Reference to substance-causes does not provide causal explanations of that sort—it does not explain why certain types of events occur when they occur. Clarke thinks that reference to *agents* as substance-causes can give us answers to some why-question. If, for instance, Sue's *A*-ing leads to a tragic accident, the question "Why did that happen?" can sensibly be answered with "Because of Sue." But Clarke acknowledges that this falls short of the kind of explanation that we typically expect in the case of human agency.[66] In particular, reference to the agent as a substance-cause does not explain *why* the agent did what she did—rather than something else. But it does not follow, firstly, that agents cannot be causes. Causal explanation, as Clarke points out, has many epistemic and pragmatic dimensions that are not shared by the metaphysical relation of causation, and the connection between causation and explanation 'may not have to be as straightforward' as is often assumed.[67] And secondly, it does not follow that substance-causation is worthless or unimportant insofar as human agency is concerned. Appealing to agent-causation, the integrated model goes beyond the standard-causal model not in terms of explanation, but in terms of *control*.[68] Adding a 'further power to causally influence which of the open

[66] Compare Clarke, 2003, p. 200.
[67] Ibid., p. 201.
[68] Compare Clarke, ibid., p. 200: 'substance causation is not appealed to by an agent causalist in order to address any problem concerning explanation. The appeal is aimed at solving a problem of control, […] and the issue of control is different from that of explanation'.

alternatives will be made actual' it can account for the kind of control that is required for libertarian free will and moral responsibility.[69]

Clarke thinks that control is a causal phenomenon, and he maintains that causation by the agent is or constitutes a kind of agential control. But it is not so obvious that the agent-causal power does in fact constitute an additional kind of control, and it is not obvious that the approach can help us to understand libertarian free will and moral responsibility. My objection to Clarke's model, and to the agent-causal theory in general, is that appeal to the agent—as a substance—cannot account for control. It does not only fail to account for the *kind* of control that is required for libertarian free will and moral responsibility, but it cannot account for the phenomenon of agential control at all. Given that, it will become clear that it fails thereby to account for SDO—self-determination as origination—as well.

Origination and Control

In chapter 4 I will explain how the standard-causal model can account for agential control. In order to see why the agent-causal theory fails to account for control, it will be instructive to anticipate some of the features of the standard-causal account.

On the standard-causal model, an agent's exercise of control is constituted by agent-involving event-causal processes. A necessary condition for such a process to be an exercise of control is that the effect-component is caused by some of the agent's reason-states. However, as examples involving deviant—or wayward—causal chains show, rationalisation and *mere* causation is not sufficient. Rather, the reason-states must cause the action in the right—that is, non-deviant—way. Generally, a causal chain leading from the agent's reason-states to the action is deviant if it runs through an event that undermines the agent's control over the action.[70] I shall argue that such cases can be blocked if it is required that the action is in addition *guided* by and *responsive* to the reason-states. In particular, the action must be guided by and responsive to the *contents* of the relevant reason-state. In chapter 4 I will discuss the notions of guidance and reason-responsiveness in more detail, and we will see that they are compatible with the standard-causal approach.

---

[69] Clarke, 2003, especially chapter 9

[70] In a much-discussed example (which is due to Donald Davidson, 1980) the reason-states cause a state of nervousness, which then causes the movement that is rationalised by those reason-states. I will discuss that example in chapter 4, pp. 160, and pp. 179.

Let us turn now to the agent-causal model. Clarke argues that we have to consider the issue of agential control alone—separated from the issue of the explanation—in order to see the advantage of the agent-causal approach. The problem, though, is that it is difficult to see how and why the agent-causal power constitutes a kind of control at all—let alone the kind of control required for libertarianism.

According to the reductive standard-causal model, an agent exercises control just in case her actions are caused, guided by and responsive to some of her reason-states. The theory explains what control consists in and how an agent's exercise of control is realised by event-causal processes. We can see or understand *why* that is an account of *control*. A central role in the account is played by the intentional *contents* of agent-involving reason-states. Nothing, though, plays a similar role in the agent-causal theory. In fact, nothing can play such a role, simply because substances do not have contents. Reference to guidance by and responsiveness to contents explains why the causal process in question constitutes an exercise of control, and it explains particular exercises of control—it explains why it is exercised in the *particular way* in which it is exercised (by the agent in the circumstances).

The exercise of the agent-causal power, however, is not guided by the agent's reason-states, character traits or some other *features* of the agent. It remains therefore unintelligible why the agent exercises that power in the particular way she does. The theory, as far as I can see, cannot explain why an exercise of the agent-causal power is an exercise of *control* at all—it does not have the resources to explain why, and in what sense, the agent controls her actions. We can agree with Clarke that control is a causal phenomenon, if that is supposed to mean that wherever there is control, there is causation. The converse, it seems clear, does not hold. It is certainly not the case that wherever there is causation, there is agential control. If we want to understand control as a causal phenomenon, we must specify in virtue of what further features a given causal process constitutes control—we must specify what distinguishes causal processes that constitute control from the ones that do not. The agent-causal theory fails in that respect. It does not explain why the agent-causal relation is not *merely* a causal relation—it does not explain why it constitutes control. Now, Clarke is aware of objections of that sort. In response he says that the

> […] objection loses force when we note that the substance in question is the agent and, moreover, one who has sophisticated rational mental capacities.[71]

Given those capacities, he says,

> […] it seems to me a credible claim that this individual's causing that decision partly constitutes the active control that she exercises in making that decision.[72]

That response is hardly convincing, though. The claim is that the agent's causing the action constitutes active control, because that very agent is a conscious and rational being. But a well-developed standard-causal model can account for all such 'rational mental capacities' and for their role in the production of action. Furthermore, by requiring that the contents of mental attitudes are causally relevant it explains why the mental states—and the involved dispositions and capacities—*guide* behaviour. It is just not obvious that the exercise of an *additional* causal power is also guided by those capacities, merely because it is one and the same agent who possesses both that additional agent-causal power and the rational capacities. The guidance provided by the mental and rational capacities may well be restricted to the kind of control described by the standard-causal model.

Clarke thinks it is *credible* to think that agent-causation provides the kind of active control required for free will, because the agent is, literally, an *originator* of her decisions and actions. We saw that standard-causal theories fail to account for free will, since they fail to provide an account of self-determination *as* origination—all they can account for is self-determination *without* origination. Let us grant Clarke that the agent is the originator of the action, because the agent causes it directly. The crucial question is, though, whether origination amounts to, or counts as, self-determination—for what is required it not *brute* origination but self-determination as origination.[73]

Recall that, according to the standard-causal model, the action is determined by the agent *himself*, because it is caused, in the right way, by the agent's reason-states. In accounting for self-determination in that way, we are referring to mental aspects *of*

---

[71] Clarke, 2003, p. 162.
[72] Ibid.
[73] By '*brute* origination' I mean a primitive metaphysical relation of causation that holds between the agent and an action (performed by that agent), which is distinct from all causal relations that hold between (events of) the agent's having of certain properties and actions (performed by that agent). In other words, that relation holds between actions and the substance *simpliciter*—as opposed to the substance's having or instantiating some property.

the agent. What we get is a plausible account of self-determination *without origination*.[74] But if we identify the self with the agent—construed as the human organism—we get nothing like that. We lose the features of the agent that help us to understand why the causal processes in question constitute self-determination; we get origination *without self-determination*. Self-determination cannot just be a brute relation of causal determination that holds between the self and the action—at least not if the self is construed as the human animal. Clearly, agents who govern themselves—who act autonomously—know what they are doing and they are in charge of what they are doing. Self-determination cannot be *mere* causation. It must be causation that constitutes agential *control*. But being a causal origin of something, as we have seen, is not the same as having agential control. (It seems to me that the point about control and the point about self-determination are two sides of the same coin, as it were. The agent-causal relation does not constitute control, because it is not causation that is guided by and responsive to relevant mental features of the self, and it does not constitute self-determination, because it does not constitute agential control.)

One may respond that the *integrated* agent-causal model establishes the desired synthesis of self-determination and origination. But that is not convincing. The aspects of self-determination and control are accounted for by the standard-causal component of the model. The agent-causal component provides the aspect of origination. What must be shown is that the agent-causal component constitutes agential control and self-determination *by itself*, which cannot be achieved by merely supplementing it with a standard-causal account of control and self-determination.

Origination and Moral Responsibility

Clarke concedes that the direct relevance of agent-causation to free will is neither obvious, nor certain. In connection with that, he says that is difficult to directly settle questions such as whether an agent is *able* to decide otherwise or whether it is really *up to* an agent to do otherwise, because we do not have 'sufficiently robust intuitions

[74] Note that this a rather weak conception of self-determination, which must be distinguished from the stronger account of autonomous agency, which will be outlined in chapter 4, pp. 196.

on the matter'.[75] We should, therefore, try to find another way to settle the issue. Clarke says that

> the best way that we have to address the question whether the account at issue secures the exercise of sufficient active control for free will is by turning to the things for the sake of which we value free will, and in particular to moral responsibility.[76]

Turning to moral responsibility, however, Clarke does not really offer anything new. The idea seems to be that the appeal to the agent as *originator* is more convincing in the case of moral responsibility than in the case of free will. Clarke, of course, is not alone in having this intuition. In fact, the idea that one is truly responsible for an action only if one has *initiated* it—only if one was its origin—seems to be a central motive behind the agent-causal approach. Roderick Chisholm, for instance, argued that an agent is not responsible for the things that are caused by her desires and beliefs, since she is not responsible for the desires and beliefs she 'happens to have'.[77] An agent is responsible for an action, only if the agent *herself* causes it. But an agent is not identical with her desires and beliefs. Chisholm says that 'if I am responsible for an event, then I initiate a causal chain [...].'[78] And an agent initiates a causal chain leading to an action, only if the *agent* is the uncaused cause or causal origin of it.[79] In support of this view, Clark invites us to consider the following example.[80]

Compare the two agents Sam, and his counterpart Sammy. Both Sam and Sammy have to decide whether to *A* or *B*. Sam lives in a deterministic world; how he is going to decide is causally determined by antecedent events. The world of Sammy is indeterministic and he causes the action in accordance with the integrated agent-causal view. Assume further that Sam is *not* morally responsible. Do we have reason to believe that Sammy *is* morally responsible? Clarke says that it 'strikes' him that Sammy 'may well be responsible' for the following two reasons. Firstly, the decision made by Sammy is not determined; he has alternative possibilities. And secondly,

---

[75] Ibid., p. 160.
[76] Ibid., p. 161.
[77] Compare Chisholm, 1964, p. 29.
[78] Chisholm, 1977, p. 624.
[79] Compare, for instance, van Inwagen's reading of Chisholm. He says that, according to Chisholm, 'for an act to be [agent]-causally determined is just *what it is* for that act to be such that its agent is responsible for it: produced by the agent himself, and by nothing else' (van Inwagen, 1977, p. 570).
[80] Clarke, 2003, p. 160.

> he exercised greater active control; he exercised a further power to causally influence which of the open alternatives would come about. In doing so, he was literally an originator of his decision [...] This is why [Sammy] is morally responsible for his decision [...].[81]

The first of the two reasons is irrelevant in the given context. As Clarke himself argues, we do not have to assume agent-causation in order to account for alternative possibilities, because a probabilistic standard-causal theory can satisfy AP as well. As with respect to the second reason, I do not find the appeal to origination more convincing than in the case of free will. The suggested kind of origination may account for *causal* responsibility. But when the question *who* caused something is settled, it may still be an open question whether the person in question is *morally* responsible. When we try to settle that question we are, usually, interested in the person's intentions, or, for instance in cases of negligence, whether the person paid as much attention as it can reasonably be expected. Our judgement whether an agent is morally responsible depends on such information *about* the agent, rather than on the information that *this* agent, rather than somebody else, caused the action in question—in philosophical debates on moral responsibility we usually assume that we already know *that*.[82]

Conclusion

To construe self-determination and an agent's exercise of control as a causal phenomenon works in the case of the standard-causal theory, because non-deviant causation of actions by reason-states requires guidance by and responsiveness to the contents of those states. Nothing can account, in a similar way, for control in the case of agent-causation. The *exercise* of the stipulated additional causal power remains mysterious, and it cannot be explained why having that power amounts to having an additional source of agential *control*.

The case for agent-causation is based on the assumption that causation by a substance is possible, that libertarianism about free will and moral responsibility is

[81] Ibid.

[82] Apart from that, there are at least two further problems with Clarke's approach. Firstly, the argument is circular with respect to the traditional approach, according to which accountability is grounded in free will, rather than the other way round. Clarke, though, treats being responsible as a criterion for having free will (and for having the required kind of agential control). Secondly, it seems that there is no way of finding out whether or not the person in question *agent*-caused the action, in the circumstances. But whether the conditions for a person's being morally responsible apply or not must be *knowable*, for we are, after all, dealing with a *practical* issue.

true, and that *only* agent-causal theories can account for libertarian free will and moral responsibility. I argued that agent-causation fails in that respect. It cannot account for libertarian free will and moral responsibility, because it cannot account for self-determination as origination and the associated kind of control. Given that, the case for agent-causation collapses.

Now we can see that the objection addresses the agent-causal view in general, rather than just the integrated version. Two points must be noticed here. Firstly, the case for the traditional agent-causal model adds to the case for the integrated model only the argument from libertarianism about acting for reasons. On that view, acting for reasons is choosing in the light of reasons with libertarian free will. But, given that the case from libertarian free will fails, the case from acting for reasons fails as well. Secondly, the objection says that the agent-causal model fails to account for self-determination and agential control. That, of course, is an objection against the agent-causal approach as such—it applies to both the integrated and the traditional model.

I shall close this section with a remark concerning the overall dialectic of the case against the agent-causal theory. We assumed, for the sake of the argument, that causation by a substance is possible. That assumption, as I noted, is very controversial, and usually it is motivated by the endeavour to defend an agent-causal theory. That is to say that belief in the possibility—coherence or intelligibility—of substance-causation is, usually, motivated by belief in agent-causation. The outlined arguments from libertarian free will, moral responsibility and acting for reasons are supposed to provide *independent* reason to believe in agent-causation. However, given that those arguments fail to support the agent-causal model, the troubles with the notion of substance-causation strengthen the case *against* agent-causation.

The Metaphysics of Agency

We have distinguished between three positions in the metaphysics of agency: the reductive model, the non-reductive model and volitionism. What is of great importance, in connection with that, is the metaphysical status of reasons for action. In particular, the question whether the reasons for an action are among its causes, and the closely related question whether reason-explanations of actions are causal explanations. We saw that reductive positions go hand in hand with a causal theory about acting for reasons and reason-explanation, whereas non-reductive views are

compatible with both causalism and non-causalism.[83] In the second and third chapter I will defend causalism about reasons and reason-explanations. In connection with that, I will argue against another position concerning the metaphysics of agency; namely, pluralism. According to pluralism, the endeavour to provide an account of the relation between reasons for action and the causes of bodily movements is a misguided project. Scientific explanations and reason-explanations are different in kind and entirely independent. Once the endeavour to account for the relationship between them is given up, the questions and problems entailed by the causal approach disappear. We can, then, distinguish between the following positions.

> Reductionism about agency: the ability or power to choose and act—possibly in the light of reasons—can be reduced to and explained by reference to agent-involving states and events and the causal and causally explanatory relations that hold between them.
>
> Non-reductionism about agency: the ability or power to choose and act—possibly in the light of reasons—is irreducible and primitive; in particular, it cannot be reduced to or explained by reference to agent-involving states and events and the causal and causally explanatory relations that hold between them.
>
> Volitionism: the ability or power to choose and act—possibly in the light of reasons—cannot be understood as a causal or otherwise relational phenomenon. Rather, an agent manifests that power or ability by performing spontaneous mental acts, which are *sui generis* acts.
>
> Pluralism: The attempt to account for the ability or power to choose and act—possibly in the light of reasons—is a misguided metaphysical project, which stems from the equally misguided attempt to specify the interrelations between the *space of reasons* and the *space of causes*.

In the next chapter I will say more about causalism and non-causalism about reason-explanations and about the pluralist stance. In the following last sections of this chapter I will, firstly, present objects to volitionism. Then I will turn to further versions of non-reductionism. And finally, I will briefly discuss two further approaches; namely, emergentism and the Kantian approach. I will show that the different versions of emergentism and the Kantian model fall under the introduced positions—that they fall under reductionism, non-reductionism or pluralism.

---

[83] Prima facie, reductionism about agency is also compatible with non-causalism about reasons. But it is difficult to see why anyone would endorse and defend this position. I will say more on that in chapter 2, pp. 108.

Volitionism

Volitionism says that all actions issue from certain basic mental acts—from volitions or acts of the will. Construed in that broad sense, the view is compatible with the reductive standard-causal and with the non-reductive agent-causal model of agency, since it may construe volitions either as being caused by agent-involving reason-states or as being caused by the agent. However, I take volitionism to be the view that volitions are not caused at all; neither by events nor by substances.[84] Further, volitions *confer* agency on other mental events and on bodily movement. That is, they are *voluntary* mental and overt *actions* just in case they are caused by volitions.[85]

I said that volitionism is widely rejected, since it is subject to a vicious regress argument.[86] That argument goes roughly as follows. According to volitionism, bodily movements and mental events are or constitute voluntary actions only if they are caused by agent-involving volitions. What about the volitions themselves? By definition, they are volitional acts themselves—otherwise they could not ground voluntary action. But if they are voluntary acts, they must themselves be caused by a volition, and so on—*ad infinitum.*

Even opponents of the view, though, have pointed out that the regress argument is by no means decisive.[87] The argument assumes that *all* voluntary actions must be caused by volitions. But that is not what volitionism says. Volitionism says that only those voluntary actions, which are not themselves volitions, must be caused by volitions. It divides voluntary action into two sub-categories: actions, which are voluntary actions in virtue of being caused by volitions, and volitions, which are voluntary actions in virtue of *intrinsic* properties. The regress argument, however, imposes the assumption that all voluntary actions are voluntary actions in virtue of causal—that is, extrinsic—properties. Given that, the objection seems plainly to beg the question.

However, things are not so straightforward; neither for nor against volitionism. I agree that the regress argument alone is not decisive. Furthermore, I think that there is no single objection to volitionism that is by itself decisive. But there are a number of

---

[84] That is why Clarke, 2003, calls it a 'noncausal' and O'Conner, 2000, the 'simple indeterminist' view.
[85] Recent defenders of the view include Hornsby, 1980, McCann, 1998, and Ginet, 1990.
[86] That argument is usually credited to Gilbert Ryle, 1949.
[87] Compare Brand, 1984; Enç, 2003.

further points, which jointly provide good reason to reject the view, and which help to restore the force of the regress argument—at least to some extent.

Firstly, to require an extrinsic or relational account of agency is not simply begging the question. Only a relational characterisation allows us to *reduce* actions to events, and there are independent reasons why such a reduction is worth wanting. To be sure, a position should not be rejected just because it is non-reductive, or just because it commits us to an additional ontological category—a category of *sui generis* acts. Generally, reductive accounts provide a better explanation of phenomena of a certain kind by reducing them to—and thereby explaining their relationship with—phenomena of a different kind. In order to see why that is the case with respect to agency, consider the following two claims. Firstly, it is very plausible to assume that both mental and overt actions are realised by events (say, by brain events and bodily movements). Secondly, it is widely held among contemporary philosophers that the order of physical events—generally, the order of events as opposed to *sui generis* acts—is causally closed. Given those two assumptions, it seems that a reductive account of agency is worth wanting, because it locates agency within the order of events—it shows us how actions relate to events by showing how agency can be realised by event-causal processes.[88] A non-reductive account faces the burden of explaining how *sui generis* acts relate to events, which constitute or realise them, and of explaining how such acts can *interact* with events without violating the closure of event-causation. No non-reductive theory of agency, as far as I know, has yet provided satisfying answers to those questions. Given that, we have good reason to prefer reductive accounts.[89]

Secondly, an extrinsic or relational characterisation is to be preferred for the further reason that intrinsic characterisations are implausible and problematic. Some philosophers have described volitions as *efforts* of the will or as *tryings*. But that seems misguided—at least as far as our common sense concepts of trying and making

[88] I will say more the on that in chapter 2 and 3.

[89] Jonathan Lowe, 1993, has suggested an emergentist account of mental causation according to which mental acts are causally efficacious not as events, but as *enabling* causes; they *coordinate* rather than initiate causal chains. Lowe argues that enabling causes can be causally efficacious without being reduced to physical events and without violating the causal closure of the physical. But that proposal misses the very nature of volitions. Recall that volitions are a manifestation of the agent's ability to perform acts spontaneously: they are mental actions by virtue of which the agent can *initiate* actions spontaneously. To construe them as enabling causes—and deny their power to initiate—cannot rescue their efficacy as volitions.

an effort are concerned. Typically, one tries to—or makes an effort to—bring something about, which involves the performance of an overt action. Playing basketball, for instance, one tries to make a basket by doing it—by throwing the ball. It seems plainly wrong to say that, in such cases, the effort or the trying is the mental component that remains when the bodily movement is subtracted.[90]

According to Ginet, the feature that distinguishes volitions from events is what he calls an '*actish* phenomenal quality', and he says that it is part of this 'I-directly-make-it-happen phenomenal quality of my mental act that I determine that it occurs precisely then, when it does.'[91] It has been pointed out, though, that this approach is hopelessly subjective.[92] That point becomes most obvious when one considers the notion of agential control. According to Ginet, an agent exercises control just in case it *seems to* the agent that she exercises control. But introspective judgements can be false. And if they are correct, there should be something *in virtue of which* they are correct. Introspection may reliably track or indicate control, but, clearly, we do not want to say that the fact that something seems or feels like being in control *constitutes* control. What further counts against volitionism, in connection with that, is that the theory does not provide an account of control—it merely provides a description of how it seems or feels like to be in control.

Thirdly, volitionism is problematic, because it is committed to a non-causal theory of reason-explanation. It denies that desires and beliefs are among the causes of both mental and overt actions, and it is therefore not compatible with the causal theory of reason-explanation. Given that non-causal alternatives are problematic, as I will argue in chapter 2, volitionism is problematic as well.

Finally, volitionism misconstrues agency. According to the view, volitions are more basic or fundamental than overt actions in the sense that the latter are actions in virtue of being caused by the former. Let us assume that mental and overt actions are *not* on a par: one of the two kinds of action is more basic than the other one. There is a strong *prima facie* case for claim that *overt* actions are in fact more basic than mental acts. What is, among other things, central to the concept of human agency is

[90] To put that point in mathematical terms is of course inspired by Wittgenstein's question: 'What is left over if I subtract the fact that my arm goes up from the fact that I raised my arm?' (Wittgenstein, 1953, 622).
[91] Ginet, 1990, p. 14.
[92] Compare O'Connor, 2000, pp. 25-26; Clarke, 2003, pp. 19-21 and Enç, 2003, pp. 18-19.

the notion of goal-directedness and the notion of doing something *for* reasons. Agents perform actions in order to achieve or attain some of their goals, and those actions are subject to explanations in terms of reasons. Now, hardly ever, if ever at all, one's end is to make up one's mind—to decide or will something. And hardly ever, if ever at all, we have reasons to decide or will something, which are not in the first place reasons to bring something about (by performing an overt action). Our goals and ends are, typically, things that can be achieved or attained only by performing overt actions, and our reasons for action are, typically, reasons for overt actions. That is, of course, not to say that we make decisions for no reason. Rather, the reasons for which we decide are, typically, the reasons for which we perform overt actions. Decisions are, typically, instrumental in the sense that we must make decisions in order to move from states of indecision to states in which we are settled on pursuing one course of action.[93] Those observations support the claim that overt rather than mental actions form the more basic or fundamental kind of action.

None of those objections is by itself decisive against volitionism. But taking all of them together gives us good reasons to reject volitionism. At the very least, it gives us reason to prefer alternative accounts that do better with respect to all or some of those points.

The Non-Reducible Self

I rejected the agent-causal theory, which is a *version* of the non-reductive approach to agency. Now I shall explain why the presented argument applies to non-reductionism in general. Most philosophers who reject the agent-causal theory do so because they reject the notion of causation involved—they reject agent-causation because they think that causation by a substance is impossible. Others are dissatisfied with the way in which the agent-causal theory construes the agent. They reject the view that the human agent is the human animal or organism.[94] So, the former object to the idea that

[93] Compare, for instance, Mele, 1992, chapter 9.

[94] As already noted, according O'Connor it is the 'official' agent-causal view that the agent is the human organism. It might be questioned, though, whether that is possible at all. One may think that the human agent is essentially a person, and that a person is essentially a self-conscious, rational and responsible agent. Given that, the human animal is certainly not essentially a human agent, because the human animal can survive change that the human agent cannot survive. The human animal and the human agent, in other words, have different persistence conditions. If one thinks that different kinds of substances are distinguished by their persistence conditions, then it seems that the human agent cannot possibly be identical with the human animal (compare for instance Lowe, 1996 and 2003b). *Prima*

actions are *caused* by substances, and the latter reject the idea that human actions are caused by the human *organism*.

The non-reductive approach says, generally, that whenever a human agent performs an action—possibly in the light of reasons—the agent stands in a metaphysically irreducible relation with that action (or with an event that is or constitutes that action). One may, then, reject agent-causation *and* endorse the non-reductive approach by denying that the irreducible relation in question is causal in nature. Or one may endorse a different version of agent-causation that does not identify the agent with the organism.

But if the relation is not causal, of what kind is it? And if the self is not the organism, what kind of thing is it? A possible answer to the latter question is to say that, since we are concerned with intentional, rational and possibly free agency, the agent must be an entity that is psychological or mental in kind. With respect to the former, one may stipulate a primitive and irreducible relation—say the relation of *performance*. We obtain, then, three further versions of non-reductionism about agency. The first says that *mental* substances irreducibly cause actions. The second says that human organisms irreducibly *perform* actions, and the third version says that mental substances irreducibly perform actions. The first option has recently been defended by Jonathan Lowe and William Hasker.[95] John Searle has suggested a position that is in line with the third version.[96] The second option is not currently held, as far as I know.

Let me now explain why my argument against agent-causation can be generalised to an argument against non-reductionism about agency. That argument says, basically, that appeal to agent-causation does not account for self-determination as origination and for the associated kind of control that is required for libertarianism. I

---

facie, it is not open to proponents of agent-causation to identify the agent with an individual that is an agent in virtue of having the right accidental properties. Causation by the agent is causation by a substance. The causal work is done, as it were, by the substance, rather than by the substance's having of a certain property—which would be tantamount to event-causation. The outlined argument, though, suggests that the human animal cannot be *that* substance. The agent must therefore be identical with a substance distinct from the human animal—say, a *psychological* rather than a biological substance (Lowe, 2003b, argues that this view does not entail that persons are *immaterial* substances). I shall not pursue this line of argument. Rather, I will assume, for the sake of the argument, that the official view, according to which the agent is the human animal, is coherent.

[95] See Lowe, 2003b, and Hasker 1999.

[96] Searle, 2001, thinks that we must postulate an irreducible 'self that combines the capacities of rationality and agency', that decides and performs actions on the basis of reasons and that is the 'locus of responsibility' (p. 95 and p. 89).

argued, furthermore, that it does not provide an account of control at all. Rather, it simply maintains that the irreducible agent-causal power constitutes control. The standard-causal approach can account for control. According to that theory, the agent exercises control by virtue of being guided by and responsive to the contents of the relevant mental states and events. The fact that the *contents* of mental states and events play a causal role is crucial to that account. Nothing, though, plays a similar or comparable role in the agent-causal theory. The theory, I argued, does not have the resources to tell us why and in what sense the agent controls or guides her actions. It is quite obvious, I think, that the alternative versions of the non-reductive view fail as well in that respect. It does not matter whether one refers to a biological or mental substance, for what is responsible for the failure is reference to a substance *simpliciter*.

In response one may say that mental substances, as opposed to biological substances, have mental properties that can explain the required kind of control. However, if it is a mental property that can be associated with a mental attitude that is of the same kind as familiar propositional attitudes—such as beliefs, desires and intentions—then the theory collapses into the reductive approach. For then there is no obvious reason to deny that it is the substance's having or instantiating that property—an event—that does the causal work. Furthermore, in chapter 4 I will argue that control constituted by mental attitudes of that kind is not the kind of control that is required for libertarianism.[97]

If, however, that mental property is supposed to constitute the having of a mental attitude of a different kind, then we would need to know more about it. Is that mental attitude a propositional attitude? Is it a cognitive or a conative attitude? Does it have a so-called direction of fit? In what relationships does it stand with the familiar propositional attitudes—what is its functional role? Can it explain the required kind of control? There are, as far as I can see, no mental attitudes for which questions of that kind can be answered and which do not fall under one of the familiar types of mental attitudes. But assume, for the sake of the argument, that mental substances have mental properties that can explain the exercise of the required kind of control. Even then, any particular exercise of that kind of control could as well be explained by referring to the substance's having or instantiating that property. And, again, there

[97] See chapter 4, pp. 198.

would be no obvious reason to deny that it is the substance-involving event, rather than the substance, that does the causal work—the theory would, again, collapse into the reductive account.[98]

Kantian Psychology

Kantian and neo-Kantian moral theories feature distinctive accounts of the nature of practical reasoning, reasons for action, personal autonomy, and moral responsibility. I do not think, though, that they provide an alternative position in the metaphysics of agency. Rather, different versions or interpretations of the broadly Kantian moral psychology can be subsumed under reductionism, non-reductionism or pluralism about agency. I shall restrict my considerations to the claims that are central to Christine Korsgaard's neo-Kantian theory and to some aspects of Kant's own position.

What is central to Korsgaard's position—central for our purposes, at least—is the role of *reflection* in the performance of action. Korsgaard says that desires and inclinations do not lead directly to action, *if* the agent is acting for a reason. When acting for reasons, agents *endorse* given desires *as* reasons. That means that the agent reflects on the normative status of what is desired, and endorses it as a reason—or as giving her a reason—by deciding to act on it.[99] Further, by endorsing a desire as a reason, the agent adopts or applies a *principle of choice*—a principle that says that incentives of that kind constitute reasons for action, in the circumstances. And by acting in accordance with it, as Korsgaard argues, the agent identifies herself with that principle of choice.

> [When you deliberate and when you act for a reason] it is as if there were something over and above all of your desires, something which is *you*, and which *chooses* which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of *yourself*.[100]

---

[98] Clarke speculates that there might be, what he calls, an 'agent-causal property' that confers an irreducible power directly onto the substance, such that the exercise of that power can only be attributed to the substance, rather than the substance's possession of that property (Clarke, 2003, p. 145). Something similar might be maintained with respect to a mental substance's performing or causing an action. But still, as long as we are not being told more about that special property, we cannot even assess whether or not the non-reductive view can thereby *explain* control.

[99] Korsgaard, 1996b, pp. 94-97.

[100] Korsgaard, 1996b, p. 100.

> [At] the moment of action I must identify myself with my principle of choice if I am to regard myself as the *agent* of the action at all.[101]

Does such a neo-Kantian position constitute an alternative position in the metaphysics of agency? According to the view, the agent is able to choose to act on certain motives in accordance with—or in the light of—principles of choice. If we want to know more about the role of the agent, we have to ask whether or not the agent's ability or power to choose in accordance with principles can be reduced to the role of the motives and principles involved. Either the agent's role can be understood in terms of—and reduced to—the role of the motives and principles in question, or it cannot. If it can be reduced, we obtain a version of the reductive view, and if it cannot, the view is a version of the non-reductive approach.

In the former case, it may well be that principles of choice assume a special role in comparison to other mental attitudes, since they are, as Korsgaard says, expressive of oneself. But there is no reason to think that their contribution in the production of action is different in *kind* than the one from other mental attitudes (such as desires, beliefs, and intentions). In particular, if the influence of mental states and events is, in general, construed as causal, then there is no reason to think that the influence of principles of choice is non-causal.[102]

The only way to avoid the disjunction that the theory falls under either the reductive or the non-reductive approach is to deny both that the influence of mental states and events is causal in kind *and* that agents have an irreducible power to act in the light of reasons. But, as will become clear in the next chapter, such a non-causal position is committed to pluralism. That shows, then, that Korsgaard's neo-Kantian model of agency falls under reductionism, non-reductionism or pluralism about agency. I shall close this section with a remark on Kant's own theory of agency.

A purported advantage of neo-Kantian theories is that they avoid all the problems that stem from Kant's doctrine of the two *worlds*: the noumenal and the phenomenal world. Very roughly, every person has a noumenal and a phenomenal self in virtue of participating in—or being a member of—both worlds. *Phenomena* are ruled by causal necessity and by the laws of nature, whereas *noumena* are free from them; they are, rather, ruled by the norms of reason and morality. Christine Korsgaard suggested

[101] Ibid., p. 241.
[102] Arguably, principles of choice are beliefs with certain contents; namely beliefs about what considerations or motives count as—or should be treated as—reasons for action.

replacing this doctrine by a theory of two *standpoints*.[103] From a theoretical standpoint we are, as all members of the natural world, subject to causal forces and our behaviour admits of scientific explanations in causal terms. From a practical standpoint, however, we can—in fact, we have to—regard ourselves as the free, rational and truly responsible sources of our own actions. Certainly, the talk about different standpoints or perspectives sounds less problematic than talk about two worlds. And it may well be that neo-Kantian positions can avoid difficult metaphysical questions and problems by adopting that move. But there is one problematic feature of the doctrine of the two worlds that a theory of two standpoints cannot avoid. Commenting on Kant, Korsgaard says that

> [to] ask how freedom and determinism are related is to inquire into the relation between the noumenal and the phenomenal worlds, a relation about which it is in principle impossible to know anything.[104]

But something similar can be said about the two standpoints as well. In fact, Korsgaard says that the theoretical and the practical standpoint 'seem strangely incongruent', and that these 'two enterprises […] are mutually exclusive'.[105] Now, all that *suggests* that both the Kantian and the neo-Kantian position are best interpreted as versions of pluralism—a position which I shall reject in the next chapter.

Emergentism

In this final section I will set out how emergentism relates to the introduced positions in the metaphysics of agency. According to John Searle, we can distinguish between

---

[103] My aim, obviously, is not to engage in Kant exegesis. But it must be noted that Kant's doctrine can plausibly be interpreted as a theory of two standpoints rather than, ontologically, of two worlds. Korsgaard acknowledges that (compare Korsgaard, 1996a, pp. 200-205), and there is textual evidence to support this reading. In the *Groundwork*, for instance, Kant says that the 'concept of a world of understanding is thus only a *standpoint* that reason sees itself constrained to take outside appearances *in order to think of itself as practical*, as would not be possible if the influences of sensibility were determining for the human being' (Kant, 1997, p. 62).

[104] Korsgaard, 1996a, p. 203. For textual evidence compare the section 'On The Extreme Boundaries of All Practical Reason' in Kant's *Groundwork*. Kant says that reason would 'overstep all its bounds' if it wanted to explain how *pure reason* can be *practical*—how, in other words, the noumenal self interacts with the phenomenal world (Kant, 1997, p. 62).

[105] Korsgaard, 1996a, pp. 204-205. Korsgaard says that the difference between an ontological doctrine of two worlds and the theory of two standpoints is that according to the former the question of the relationship between the worlds cannot be *answered*, whereas on the latter that question cannot be coherently *asked*. In the next chapter I will turn to the closely related question of how reasons for actions relate to the causes of bodily movements, and it will became clear that this question can coherently be asked.

the following two senses of *emergent*.[106] The features of some systems can be deduced from—figured out or calculated, as Searle says—from the features of the entities that compose the system. Searle mentions shape, weight and velocity as examples. The features of other systems, though, cannot be explained in that way. Rather, in order to explain them we have to take into consideration not only the features of its constituents, but also the causal relations in which they stand. In other words, the features of such systems must be explained in terms of the intrinsic and extrinsic properties of the entities that compose it. Searle calls system features of that kind 'causally *emergent* system features'.[107] Examples are liquidity and solidity, and Searle maintains that consciousness is an emergent property in that sense.

According to a second, 'much more adventurous conception', as Searle says, an entity is emergent just in case it has—or confers—causal powers that cannot be explained by the causal interactions between its components. That there is emergence in the first sense seems entirely uncontroversial—most philosophers would not even call it *emergence*. I shall therefore focus solely on the second conception. What is central to emergence, in that sense, is the notion of irreducible or emergent causal powers.

Consider a system S that is composed of the entities $a_1, \ldots a_n$. S is a higher-level—or macro—entity that has higher-level properties, and the a_i are lower-level—or micro—entities that have lower-level properties. Suppose that S has, or instantiates, the property P. Then, P is an *emergent* property of S just in case the a_i are caused to behave differently in virtue of S's instantiating P—differently than they would have behaved, if S had not instantiated P. In that case, S has irreducible or emergent causal powers in virtue of instantiating P. Since the cause of the relevant causal relations is located at the higher-level, and since the effect is located at the lower-level, it is common to call that kind of causation *downward* causation.[108]

That characterisation is about emergent *properties*. There are, however, versions of emergentism according to which the emerging entity is not a property but an

[106] Searle, 1992, p. 111.
[107] Ibid., my emphasis.
[108] There is, of course, disagreement concerning the details of emergentism. In particular, it is not clear whether an emergent property changes the behaviour of some lower-level entities by conferring emergent causal powers onto the *lower*-level entities (and by 'superseding' the lower-level laws), or whether it confers the emergent causal powers onto the *higher*-level entity, which then literally causes the lower-level entities to behave differently. Compare Hasker, 1999, especially pp.174-177 and Clarke, 2003, pp. 177-178 and the appendix to chapter 10.

individual—an object or substance. Further, we can distinguish between two versions of property emergentism. We obtain, then, the following three—relevant—versions of emergentism. Each version, as we will see, falls under either reductionism or non-reductionism about agency.

According to the first version, there are emergent *properties* the instantiations of which constitute an agent's having or being in a familiar mental attitude, such as believing, intending, feeling, hoping, and so forth. This version of emergentism is compatible with reductionism about agency, since it concerns only the causal powers of agent-involving mental states and events. It makes no claims about irreducible powers of the *agent*, since it is about the causal powers of the agent only indirectly by being about the causal powers of agent-involving states and events. That kind of emergentism received considerable attention in the so-called mental causation debate. Some philosophers have argued that non-reductive theories of the mind[109] are committed to emergent causal powers and downward causation.[110] But since that concerns only the causal properties of mental states and events, rather than the agents themselves, that first version of emergentism falls under reductionism about agency.

According to a second version, there are emergent *properties* in virtue of which an *agent* has downward agent-causal powers. Such properties, it seems obvious, cannot be associated with familiar types of mental attitudes and events. This version is about irreducible and emergent causal powers of the agent, rather than of agent-involving states and events, and it falls therefore clearly under non-reductionism about agency. (Note that this version is compatible with both the view that the agent is the human organism and the view that the agent is a mental substance).

According to the third version, what is emergent is the higher-level *individual* or *substance*, rather than some property. What emerges, in other words, is the agent—the person or the self.[111] On that view it is not only the case that the higher-level individual has emergent downward causal powers in virtue of having emergent properties. But it is the individual itself that is emergent—its mode of existence is being emergent from, rather than being composed by lower-level entities. Given that

[109] Reductive and non-reductive theories of the *mind* are, first and foremost, about agent-involving mental states and events, and they must be strictly distinguished from reductive and non-reductive theories of *agency*.
[110] Kim, most prominently, has argued along that line. See for instance essay 17 in Kim, 1993.
[111] Compare Hasker, 1999; especially, p. 194.

this view also recognises the human organism as a substance, it is a kind of substance *dualism*. William Hasker describes it as follows.

> Emergent dualism […] recognises that a great many mental processes are *irreducibly* teleological and cannot be explained by [brain processes]. [The] power attributed to matter by emergent dualism amounts to this: when suitably configured, it generates a field of consciousness that is able to function teleologically and exercise libertarian free will, and *the field of consciousness in turn modifies and directs the functioning of the physical brain*.[112]

From what has been said so far, and from that passage, it is clear that the third version of emergentism falls under non-reductionism about agency.

Emergentism, then, is not an alternative approach in the metaphysics of agency. Rather, by specifying the relations between higher- and lower-level properties and causal powers, it provides ways to explain how reductionism or non-reductionism about agency *works*. It must be noted that it is not obvious whether or not non-reductionism is committed to emergentism. There is disagreement over the question whether the theory of agent-causation requires that agents have emergent downward causal powers and whether it requires the so-called *supersession* of lower-level laws. Most proponents of agent-causation think that the theory is committed to emergent downward causation and supersession of laws.[113] Clarke, however, argues that it is not committed to emergentism in that sense.[114] Further, the relevant kind of emergentism is essentially a view about causal properties, causal powers and the special kind of downward causation. But we saw that proponents of non-reductionism about agency may deny that an agent's irreducible powers are *causal* powers. Finally, note that it is also not obvious that reductionism about agency is committed to downward causation or irreducible causal powers (of agent-involving mental states and events). In fact, in chapter 3 I will defend a version of the reductionism about agency that is not committed to emergentism in that sense.

[112] Ibid., p. 195.
[113] Compare O'Connor, 2000; Hasker, 1999; and Dupré, 2001.
[114] Clarke, 2003, pp. 177-181

Chapter Two: Reasons and Causes

To decide whether or not reasons are causes is of central importance in the philosophy of action and in the philosophy of mind. Approaching that question, it is important to distinguish between the extensional relation of causation and the intensional relation of causal explanation.[1] Accordingly, we can distinguish between the following two claims concerning the question whether or not reasons are causes. Firstly, there is the claim that the reasons for which an agent performs a certain action are among the causes of that action, and secondly, there is the claim that explanations of actions in terms of the agent's reasons are causal explanations. On the face of it, those two claims are saying the same. Or, if they do not claim the same, it seems that they stand and fall together. In the first part of this chapter I will consider some ways in which the two claims may come apart. I will defend both claims against objections and I will argue that there is no reason to think that they do not stand and fall together. In the second part, I shall present arguments for causalism about reasons and reason-explanation and I will argue against non-causalism on the ground that it fails to provide a satisfactory account of the metaphysics of agency.

Two Kinds of Causalism, Reasons and Mental Attitudes

Let us first distinguish between two kinds of causal theories—two kinds of causalism—that correspond to the two claims introduced. The first claim says that the reasons for which an agent performs a certain action are among the causes of that action. That is a claim about the metaphysics of agency, as it concerns the metaphysical relation between actions and reasons—it concerns the efficacy of reasons and the causal history of actions. I endorsed and defended that kind of causalism—call it causalism about reasons—in the previous chapter by defending the reductive standard-causal model of agency.

The second claim concerns the nature of reason-explanations. When we give a reason-explanation of an action we explain the performance of that actions in terms of some of the agent's mental states and events that rationalise its performance. Consider, for instance, an ordinary action such as Sue's opening her handbag. Sue

[1] Compare, for instance, Davidson, 1980, essay 7 and Strawson, 1985.

wants to get her keys in order to unlock the door, and she thinks that the keys are in her handbag. Why did Sue open her handbag? She opened it because she intended to get her keys and because she believed that she would find them in her handbag. And she intended to get her keys, because she wanted to unlock the door. Sue opened the handbag for those reasons—because she had the mentioned intention, belief and desire. The performance of the action can be rationalised in the light of those mental attitudes. The second kind of causalism—call it causalism about reason-explanations—says that rationalising explanations of actions in terms of the agent's reasons are causal explanations, which is to say that the explanatory force of reason-explanations is partly causal.

Given the distinction between the extensional relation of causation and the intensional relation of explanation, it would be wrong to think that causalism about reasons and causalism about reason-explanations are saying the same. Nevertheless, it seems clear that they stand and fall together. In the previous chapter we saw that action can be defined in two different ways. It can be characterised as goal-directed and motivated activity that admits of rationalising explanation, or it can be characterised in terms of its causal history. We saw that, according to the standard-causal model of agency, the two characterisations are equivalent: to provide a rationalising explanation of an action as goal-directed and motivated activity *is* to refer to mental states and events that played a causal role in the performance of the action. That means that the standard-causal model and causalism about reason-explanations stand and fall together, and since causalism about reasons is part of the standard-causal model, causalism about reasons and causalism about reason-explanations stand and fall together as well.

A further reason to think that the two claims—and the two kinds of causalism—must stand and fall together is provided by the following simple argument. To say that a reason-explanation is a causal explanation is to say its *explanans* causally explains the *explanandum*. A necessary condition for that is that the *explanans* refers to a cause—or to causes—of what is explained. The *explanantia* of reason-explanations are the agent's reasons for the action. Therefore, in order to explain actions *causally*, reasons must be among the causes of those actions. Consider Sue's opening her handbag. A causal explanation of that action in terms of Sue's

desires, beliefs and intentions is true only if those mental attitudes were among the causes of the action.

Given that, it seems clear that the two kinds of causalism stand and fall together, and that the standard-causal model of agency goes hand in hand with causalism about reason-explanations. In the following, I shall consider three positions on related issues on the basis of which one may argue that the two claims can come apart. Given the apparent connection between them, the result that the two claims do not stand and fall together would be of considerable significance. In particular, it would open the possibility of *mixed* views—of positions that are causal with respect to either the metaphysics of agency or reason-explanation, but not both. Further, those three positions are of interest, since they give raise to three different objections to causalism.

Anomalous Monism

According to a traditional model of causation—the covering law model of causation—, an event-token of type *C* causes an event-token of type *E* only if there is causal law that relates instantiations of *C* and *E* as cause and effect. In other words, the two event-tokens must instantiate a law by instantiating types that are, as a matter of nomological necessity, related as cause and effect. If that is the case, the two events are *covered* by a law. According to one version of the covering law model, only strict generalisations count as laws, and only strict laws can ground singular causal claims.[2]

The covering law model is a theory of causation and causal explanation. It says that an event is a cause and causally explanatory only if it is covered by a strict law. It is commonly accepted, though, that there are no *strict* intentional laws that cover reasons and actions.[3] Assuming that the covering law model holds for causation and causal explanation in general, it follows that reasons are not causes and that reason-

[2] That view has been introduced as the covering law model by Hempel and Oppenheim, 1948. The underlying regularity view, of course, goes back to Hume, 1748.

[3] By *intentional* laws I mean laws that are formulated in psychological or intentional terms. Reason-explanations are rationalising explanations. In order to rationalise the performance of an action in terms of reasons, both the action and the reason must be described in intentional or mental terms. Hence, generalisations that cover reasons and actions have to refer to reasons and actions under their intentional descriptions. So, intentional laws are laws that refer to reasons and actions under their intentional descriptions. This point is of importance, if it is assumed that reasons and action have both intentional and non-intentional descriptions. If it assumed, for instance, that overt actions are token-identical with bodily movements, then reason-explanations must refer to it as an action, rather than as a bodily movement. Compare, for instance, Stoutland, 1986.

explanations are not causal explanations. Donald Davidson, however, showed that one can coherently accept that there are no strict intentional causal laws connecting reasons and actions *and* deny that conclusion—he showed, in other words, that anomalism about the mental neither entails that reasons cannot be causes, nor that reasons cannot causally explain actions.[4]

Central to Davidson's argument is the aforementioned distinction between the extensional relation of causation and the intensional relation of causal explanation. The former is a metaphysical relation and holds between particular events or event-tokens, whereas the latter explanatory relation holds between descriptions of those events or event-types. Davidson endorsed the Humean claim that every true singular causal claim entails a covering law, but he also pointed out its ambiguity.

> [The Humean claim] may mean that 'A caused B' entails some particular law involving the predicates used in the description 'A' and 'B', or it may mean that 'A caused B' entails that there exists a causal law instantiated by some true description of A and B.[5]

According to Davidson, the second interpretation is the correct one. That is, for two events to stand in the relation of causation, it is required that there is only *some* description of them under which they instantiate a causal law.

Suppose, then, that an agent's *A*-ing can be explained by referring to the agent's having the reason of type *R*. Suppose further that the intentional description of the action refers to the event *e*, and that the intentional description of the mental event *R* refers to the event *c*. According to Davidson's interpretation of the Humean claim, *c* can be the cause of *e* even if there is no law that relates them as cause and effect under their intentional descriptions. All that is required is that there is some description of *c* and *e*, respectively, bringing them under a covering law.[6] That is sufficient to show that reasons can be causes of actions, even if there are no intentional laws—laws that cover them as reasons and actions.

According to Davidson, reasons are causes and they causally explain actions. One may, however, use parts of Davidson's position in order to argue in the following way that reasons fail to *explain* actions causally. What Davidson's position—which is

---

[4] Compare Davidson, 1980, essay 11 and 12.
[5] Davidson, 1980, p. 16.
[6] Davidson thought that strict laws are only to be had in the physical sciences. Hence, what is required, according to Davidson, is not just some description, but a *physical* description of both *c* and *e* that instantiates a strict law.

known as *anomalous monism*—shows is that mental events can be the causes of actions, even if, firstly, mental event-types are not identical with physical event-types, secondly, even if there are no psychophysical laws that correlate mental and physical events, and thirdly, even if there are no laws covering the events under their intentional descriptions.[7] However, for a mental event-token and an act-token to be related as cause and effect, there must be some causal law that covers them. Given the claims of anomalous monism, that must be a non-intentional law. To say, however, that there are no causal laws covering the events under their intentional descriptions is just to say that there are no causal laws covering them *as* reasons and actions. And that just means, so the argument goes, that reasons do not causally *explain* actions. Rather, descriptions of reasons merely refer to the event-tokens that are the causes of actions.

That argument, if correct, shows not only that the two kinds of causalism do not stand and fall together. But it also provides a straightforward objection to causalism about reason-explanations. In a recent book, G. F. Schueler has advanced such an argument against causalism in his defence of a teleological account of reason-explanation. Since causalism, as Schueler argues, cannot establish that reasons causally explain behaviour it fails to provide an account of the 'explanatory force' of reason-explanations. And since one of the prime objectives of the causal approach is to provide an account of the explanatory force of reason-explanation, causalism fails in its own terms.[8]

One important assumption of that argument says that there are no strict intentional laws. That assumption, as already mentioned, is generally accepted. Another assumption says, roughly, that reference to an event under a certain description causally explains another event under a certain description only if there is a strict causal law that covers the events under those descriptions. In our case, the assumption is that reasons causally explain actions, only if there are *strict* laws that cover the events in question under their intentional descriptions—*as* reasons and actions. That assumption, however, is very controversial.

---

[7] The position is a form of *anomalism* in the sense that it denies the existence of psychological laws and in the sense that it denies the existence of psychophysical bridge-laws. It is a form of *monism* in the sense that it assumes token-identity: mental event-tokens are identical with physical event-tokens. Compare Davidson, 1980, essay 11 and 12.

[8] Compare Schueler, 2003, especially pp. 8-12.

Already in his 'Actions, Reasons, and Causes', Davidson argued that it is an 'error to think that no explanation has been given until a law has been produced'.[9] All we need to notice is that many ordinary explanations, which are clearly causally explanatory, are not grounded in strict causal laws. Davidson's example is a catastrophe that can be explained simply by citing the event of an earthquake as its cause—and it would, as Davidson says, be ridiculous to think that there are strict physical laws that connect hurricanes and catastrophes. Countless further examples of that sort could be put forward, which, I think, shifts the burden of argument to the opponent. Given that point concerning common causal explanation and knowledge, the assumption that causal explanations presuppose strict causal laws is in need of independent justification.

Others have argued that it is unreasonable to require *strict* causal laws in the case of intentional explanation, given that many scientific laws—typically the laws of the so-called special sciences—hold only *ceteris paribus*. Louise Antony, for instance, argues that nothing in our actual scientific practice 'lends support to the idea that the laws backing causal claims must be strict—there's nothing inadequate or insufficient about *ceteris paribus* laws from a *pragmatic* point of view'.[10] Given that, one cannot plausibly demand that reason-explanations must be grounded in strict intentional laws.

On the basis of that, one may then provide an alternative account of causal explanation. It is common to appeal to intentional *ceteris paribus* laws, the fact that reason-explanations support counterfactuals, or the fact that intentional regularities are backed and mediated by underlying physical mechanisms. Many proposals in recent philosophy of mind develop either one of those options or they feature a combination of some or all of them.[11]

Given, then, that causal reason-explanation does not presuppose strict intentional laws, it does not follow that reasons fail to explain actions causally, and the outlined case against causalism collapses. What opponents of causalism would need to show is

---

[9] Davidson, 1980, p. 17.

[10] Antony, 1995, p. 438. The source of that position is Fodor 1974. For a defence of *ceteris paribus* laws against the charge that they are vacuous and uninformative see Fodor 1989 and 1991, Antony 1995, and Pietroski 2000, Chapter 4.

[11] For an account of causal reason-explanation in counterfactual terms see Schiffer, 1991, Sosa 1984, Mele, 1992 and Ruben, 2003, Chapter 6. For positions that appeal to *ceteris paribus* laws and to underlying physical mechanisms see, for instance, Fodor, 1989, Segal and Sober, 1991, and Antony, 1995.

either that strict laws are necessary for causal explanation or that none of the alternative accounts—in terms of *ceteris paribus* laws, counterfactuals or underlying mechanisms—can ground causal reason-explanations. I am not aware of any convincing non-causalist argument in support of either of the two claims. We can conclude, then, that the argument from anomalous monism does not show that the two kinds of causalism can come apart, and that the outlined argument against causalism fails as well.

Externalism About Content

The argument discussed in the previous section was supposed to show that reasons do not causally explain actions, even if they are among their causes. The following position about the nature of mental content grants that reason-explanations are causal explanations. Given that position, though, one may argue that reasons cannot be causes of actions.

Let us assume that well-known twin earth thought experiments show that the contents of some mental attitudes are not determined by agent-intrinsic states only, because they are partly dependent on and determined by environmental or circumstantial states.[12] While Oscar on earth wants a glass of water, his intrinsically identical twin on twin earth, Toscar, wants a glass of twater, due to the fact that the substance that tastes and smells like water on twin earth is not H_2O. Oscar and Toscar are in different mental states, due to different mental contents, even though there is no difference in intrinsic properties between them.[13] That means that the contents of some mental states are not determined by agent-intrinsic states only, and that some mental states do not supervene on—and are not realised by—agent-intrinsic states only. In other words, some mental states are environment-dependent. That view is known as externalism about content.

Consider, then, the following argument against the claim that reason-states are causes of action. Assume, firstly, that externalism about content is true. Secondly, assume that an agent's causal powers are determined by its intrinsic properties only. (That means, roughly, that an agent's bodily movements are determined by

[12] For the notion of intrinsic property compare chapter 1, p. 15, note 18.
[13] The original arguments for that position can be found in Putnam, 1975, pp. 251-271. Compare also Burge, 1979.

instantiations of some of its intrinsic properties only—by agent-intrinsic states and events only.)[14]

According to the standard causal model, overt actions are bodily movements, which are caused by appropriate mental states and events in the right way. So, if reasons are causes of overt actions, they are causes of bodily movements. By assumption, the causes of bodily movements are agent-intrinsic states and events. In other words, if something is not an agent-intrinsic state or event, then it is not a cause of the agent's moving her body. By assumption, reason-states are not realised by agent-intrinsic states only. Therefore, neither reason-states nor their physical realisations are among the causes of bodily movements. Hence, neither reason-states nor their physical realisations are among the causes of overt actions.

Given that, one may argue against causalism about reason-explanation in the following way. It is plausible to assume that a causal explanation of an action of type *A* in terms of reason-states of type *R* presuppose that the description of the agent's being in *R* refer to a cause of the agent-involving event that is or constitutes the agent's *A*-ing. The argument from externalism about content shows that reasons are not among the causes of the relevant agent-involving events (bodily movements). Therefore, reasons do not explain actions causally.[15]

The argument from externalism, however, is a bad one. Its conclusion is ambiguous, and it follows from the assumptions only if interpreted in a certain way. Consider the following case. Suppose some agent, *S*, performs a bodily movement *m*, which is or constitutes an *A*-ing, for the reason *R*. And assume that *m* is caused by the agent-intrinsic event *n*. The argument from externalism grants that actions are bodily movements—it grants that *m*'s being caused in the right way is an *A*-ing.[16] It grants

---

[14] The idea is to treat the powers of an agent just like the powers of any other object, and the causal powers of objects are, presumably, determined by their intrinsic physical properties. Compare, for instance, Child, 1994: 'Everything we know from the sciences of matter supports the view that a thing's causal powers are exhaustively determined by its internal physical make-up' (p. 187).

[15] Can one possibly deny that reasons are causes and hold that reason-states causally explain actions? One may argue that the explanatory relevance of reasons in causal explanations of actions is grounded in, for instance, intentional *ceteris paribus* laws and the fact that reason-explanations support the right counterfactuals. And one may deny that the causal relevance of reasons is in need of further *metaphysical* vindication by grounding it in the causal efficacy of physical events (compare, for instance, Burge, 1993 and Baker, 1993). I will return to this position, which I call pluralism, later in this chapter, pp. 102.

[16] The assumption that actions *are* bodily movements is not essential to the argument, nor to the response. Similar considerations apply in case actions are construed as being *constituted* or *realised* by bodily movements. Further, nothing depends on the fact that we focus on *overt* actions and bodily

that reasons are agent-involving mental states and events, that mental states and events are realised by physical states and events, and it grants that the causal efficacy of mental events consists in the efficacy of the physical events that realise them.[17] It denies only that mental states and events are realised by agent-*intrinsic* states and events alone.

Assume further that *p*, an instantiation of the physical state *P*, realises *S*'s being in *R*. Given externalism, *p* is not an agent-intrinsic state. It follows that *p* is not identical with the cause of *m*. What follows, in other words, is that *p* is not identical with *n*, since *n* is agent-intrinsic, whereas *p* is not. From that it follows that *S*'s being in *R* is not a cause of *m*, and, therefore, *S*'s being in *R* is not a cause of *S*'s *A*-ing.

However, it is misleading to present that as a conclusion against causalism for the following reason. It is certainly possible that both *n* and *p* are complex or structured events—events that have events as their parts or constituents. Given that, it is possible that the events that constitute *n* are *among* the events that constitute *p*. And, if that is the case, it is *trivially* true that the event that realises *S*'s being in *R* is not identical with *n*, simply because the physical events that constitute *n* are among—that is, are a proper part of—the events that realise *S*'s being in *R*.

What is important to note is that causalists do *not* have to reject the principle that a causal explanation of an action of type *A* in terms of reason-states of type *R* presupposes that the description of the agent's being in *R* refers to a cause of the agent-involving event that is or constitutes the agent's *A*-ing. All they need to reject is a narrow reading of that principle, according to which the whole complex event that is the realisation of *S*'s being in *R* must be the cause of the movement, *m*, that constitutes the A-ing.

Alternatively, causalism can subscribe to an interpretation that says, roughly, that a causal explanation of *A*-ing in terms of the agent's being in *R* presupposes, firstly, that the description of the agent's being in *R* refers to the physical realisation of the agent's being in *R*, and, secondly, that this realisation is *partly* constituted by events which are among the causes of *m*—and which are, therefore, among the causes of the

---

movements. Similar considerations apply to mental actions and the agent-involving events that constitute or realise them.

[17] It grants what Kim calls the *causal inheritance principle*: 'If *M* is instantiated on a given occasion by being realised by *P*, then the causal powers of *this instance of M* are identical with (perhaps, a subset of) the causal powers of *P*' (Kim, 1993, p. 355, and compare Kim, 2000, p. 54). I will return to the problem of mental causation in chapter 3, pp. 142-153.

agent's *A*-ing. If that principle holds, then we refer—among other things—to the events that are among the causes of the bodily movement *m*, which is or constitutes the agent's *A*-ing, whenever we give a true reason-explanation of a performance of an action of type *A* in terms of *R*.[18] The opponent's argument does not show that causalism is committed to the mentioned narrow interpretation. In the absence of other arguments for the narrow reading, the causalist is justified in employing the alternative interpretation.

Are reasons, then, the causes of action? Given externalism, reasons are not the causes of action *in the sense that* the whole complex physical realisation of a reason-state is not *the* cause of the movement that is or constitutes the action. But, as we have seen, causalism does *not* depend on the truth of the claim that reasons are causes in that narrow sense. Reason-states are realised by complex physical states and events. If the causes of the relevant bodily movements are among those states and events, then descriptions of reason-states refer to the causes of the relevant movements, and they can, thereby, be causally explanatory of action. According to causalism, reasons are causes of actions *in that sense*.

Do the two claims—the claim that reasons are causes of actions and the claim that reasons causally explain actions—stand and fall together? Given the considerations of this section, we can say that they do stand and fall together insofar as causalism about reason-explanations presupposes that reasons are causes in *some* sense.

Externalism About Reasons

The following position on the nature of normative reasons for action does not directly challenge the claim that the two kinds of causalism stand and fall together. Rather, it challenges the way in which claims about reasons and reason-explanations have been interpreted. We assumed, so far, that explanations of actions in terms of reasons are explanations in terms of mental states or events. We assumed, apparently, that reasons

[18] Compare, for instance, Kim (1993, p. 304), Segal and Sober (1990, p. 19) and Noordhof (1990, p. 312) who argue that it is sufficient that physical micro-properties of the agent form only *part* of the supervenience base of mental properties. Compare also with Shoemaker's distinction between the *total* and the *core* realisation of a mental state (Shoemaker, 1981). A certain brain event, for instance, would be the core realisation of a belief state, but it would only be *part* of the total physical realisation, which involves both the brain state and certain environmental states (the brain state, in other words, is not sufficient for the agent to be in the belief state). For general discussion of externalism about content in connection with the problem of mental causation see Jackson and Petit, 1988 and Mele, 1992.

for action *are* mental attitudes. The rationale behind that assumption is, roughly, that the performance of certain actions appears as rational and intelligible in the light of mental attitudes, such as beliefs, desires and intentions, and that, when we refer to rationalising mental states, we are, therefore, referring to the agent's reasons for the action. When we were asking whether reasons are causes and whether reason-explanations are causal explanations, we were, in effect, asking whether an agent's reason-states are causes of the action and whether explanations in terms of those mental attitudes are causal explanations. Let us a call the view that an agent's reasons for actions *are* agent-involving mental states and events *internalism* about reasons for actions.[19]

The negation of that view—externalism about reasons—says that reasons are not identical with the agent's mental attitudes. Jonathan Dancy has recently developed and defended such a position. Dancy argues that the normative nature of reasons rules out that they can be psychological entities of any kind. Because our beliefs might be false and our desires inappropriate, the mere fact that we happen to believe *this*, or happen to desire *that*, does not, as Dancy argues, give us any reason, let alone *good* reason, to do anything.[20]

Let us assume that Dancy is right. Reasons are not mental states. What else may they be? Dancy discusses two suggestions. On the first, reasons are the *contents* of mental states. On the second, they are *facts* or *states of affairs*.[21] Dancy argues that neither suggestion supports the claim that reasons are causes, and he thinks that reason-explanations are, therefore, not causal explanations.

Dancy does not question that the two kinds of causalism stand and fall together. Rather, he seems to presuppose that they go hand in hand. What is important to notice is that Dancy's view on the nature of reasons entails neither that reason-states cannot

---

[19] Despite some affinities, that kind of internalism must be sharply distinguished from what is known as internalism about reasons in normative ethics—a view that is due to Bernard Williams, 1981, essay 8. According to Dancy, that view says that an agent *S* has a good reason to *A* only if, were *S* to know all the relevant facts, and deliberate rationally, *S* would be motivated to *A* (compare, Dancy, 2000, p. 15). The central idea is that whether an agent has normative reasons to *A* depends—in some sense—on which pro-attitudes (desires, motives, or values) the agent actually has. What I call internalism, however, makes no such claim about normative reasons. It requires that the performance of the action can be rationalised in the light of some of the agent's mental attitudes, but it leaves it open whether the agent actually had normative reason to perform the action.

[20] See Dancy, 2000.

[21] What reasons *really* are, on either view, depends of course on what theory of mental content, facts, or states of affairs one endorses. For present purposes the offered formulations should be good enough.

be causes of actions, nor that they cannot be causally explanatory. It does, in other words, not entail that explanations of actions in terms of mental states and events cannot be causal explanations. That is because all that Dancy's position claims and entails is about *reasons*, rather than reason-states.

Dancy is aware of that and he considers the possibility that there are two distinct intentional explanations of one and the same action; a normative and non-causal *reason*-explanation and a causal explanation in terms of the agent's mental states and events. But, eventually, Dancy rejects that possibility.[22] He argues that externalism about reasons is incompatible with both causalism about reasons and causalism about reason-explanations. Dancy's reasoning, however, is based on an overly narrow interpretation of the claims of causalism. I shall now suggest an alternative interpretation of causalism, which is compatible with externalism about reasons.

The suggestion is that we can plausibly *call* certain kinds of explanations *reason-explanations*, if their *explanantia* stand in some appropriate relation with the agent's reasons for the action. What I mean by that should become clear in due course, when I consider each of the three views concerning the nature of reasons—the view that reasons are mental attitudes, contents of mental attitudes, and the view that they are facts. What we have to assess, in each case, is whether the relation between mental attitudes and reasons is such that a causal explanation of an action in terms of mental attitudes can plausibly count as a reason-explanation.

Firstly, if one assumes internalism about reasons, the relation in question is identity. Reasons are mental states, and causal explanations of actions in terms of those mental states can therefore be called reason-explanations.

Secondly, assume that reasons are the *contents* of the agent's mental states. The relation in question is then the relation that holds between mental attitudes and their contents. I think we do not have to go into any detail concerning the nature of mental content and the relationship between attitudes and their contents, for it seems

[22] The basic structure of Dancy's argument (which can be found in Dancy, 2004) is the following *reductio ad absurdum*. Assume that there are causal explanations of actions in terms of mental states and events in addition to reason-explanations. Dancy argues that such explanations must be normative explanations. Normative explanations of actions give the agent's reasons for the action. But that cannot be right, because, firstly, mental attitudes are not reasons for action, and secondly, because the 'normative underpinning' of explanations in terms of mental attitudes is 'at odds with their supposedly causal nature' (p. 39). Dancy's argument is of course more sophisticated than that, and it draws on results from Dancy, 2000, where it is argued that the reasons for which an agent acts are not psychological states and events.

straightforward that this view supports the claim that causal explanations of actions in terms of mental states can reasonably be called reason-explanation. The view construes reasons as the contents of mental states. Hence, to cite reason-states is to cite reasons as their contents.

That view says that causal explanations in terms of mental states can be *called* reason-explanations, even though reasons are, as the contents of mental attitudes, not causes. It would be a mistake, though, to think that that shows that the two kinds of causalism can come apart. We assumed, *initially*, that reasons are mental states. This assumption has been suspended for the sake of an assessment of the present view. Without that assumption, the present view has no direct bearing on causalism. Given the initial assumption, the claim that reasons are causes *is* the claim that reason-*states* are causes and the claim that reasons causally explain actions *is* the claim that reason-*states* causally explain actions. However, if we dismiss the assumption, which claims should be regarded as distinctive of causalism? On my view, causalism says that reason-*states* cause and causally explain actions and that causal explanations in terms of reason-*states* can be called reason-explanations. On that construal, causalism is not committed to the claim that reasons are causes in its literal and strict sense (provided that reasons are not identical with reason-states).[23]

Let us now turn to the third view, which says that reasons are *facts*. Dancy arrives at the conclusion that reasons cannot be mental states by stressing their normative nature. But he acknowledges that reasons have a motivational aspect as well. In order to be explanatory of the performance of an action, reasons must have motivated the agent to perform the explained action. That motivational requirement is a necessary condition for something to be explanatory as a reason for action.

In order to be motivated by a reason, as it seems very plausible to assume, the reason must in some way play a role in the psychological process that leads the agent to the performance of the action. Even if we assume that reasons are facts, rather than mental attitudes, we must account for their motivational role in *psychological* terms.

---

[23] Note that on the present view *reasons* can nevertheless be construed as being causally explanatory, if it is granted that the contents of mental states are causally relevant, in the sense that mental states have their effects *in virtue of* their contents. Given that, one can hold that reasons are causally relevant and explanatory without being—strict speaking—causes. For defence of the claim that mental events are efficacious in virtue of their contents (or in virtue of those properties in virtue of which they possess a certain content) see Jackson and Petit, 1988; Segal and Sober 1990; Mele 1992; Braun, 1995; and Noordhof, 1999.

The obvious way to do so is to say that reasons, as facts, can motivate the agent to do whatever they recommend or favour, if the agent has—or is in—a mental state that *represents* the reason. Construed in that way, the present view is actually very close to the previous one. The difference is that reasons are not identified with the contents of mental states, but with whatever is represented by them—with the facts or states of affairs that mental states are *about*. Fortunately, that difference does not prevent us from drawing the same conclusion as in the previous case. A reason is explanatory of the performance of an action only if having the reason motivated the agent to act for it—or in accordance with it. This motivational aspect is best understood in terms of psychological states that play a role in the agent's motivational economy *and* that represent the relevant reason-facts. A successful reason-explanation, it seems, must refer to a mental state or event that motivated the agent and that represented the fact that was the agent's reason for the action.

Given that, we can see that causalism about reason-explanations is in fact compatible with externalism about reasons. According to causalism, explanations in terms of mental states and events, which render the performance of an action rational and intelligible, are causal explanations. According to externalism about reasons, reasons are facts. In order to see why they are compatible let us consider two ordinary examples in which a reason-explanation is given solely in terms of facts—without mentioning any of the agent's psychological states. "Why did you stop the car so abruptly?"—"Because there was a child crossing the street." "Why did you take that medicine?"—"Because my doctor recommended it." In cases like that we can provide a broader explanation, which mentions some of the agent's mental states. The driver stopped, because she noticed that a child was crossing the street. The patient took the medicine, because he wanted to do whatever fosters the healing process, and because he believed that the doctor gives good advise on that. Quite often, of course, we do not mention the psychological states that *mediate* between the facts and the agent's awareness of them, simply because the way in which the non-psychological explanation is formulated implies such a mediation. However, when we say that the driver stopped, because there was a child crossing the street, it is not only that everyone will *assume* a connection between that fact and the driver's response, which is psychological in kind. But we *have to* assume such a connection—otherwise the explanation would not make any sense. Reason-explanations that are given merely in

terms of facts or states of affairs are, in general, condensed versions of reason-explanations that feature mental states and events.[24] Often the condensed version is sufficiently explanatory for everyday and practical purposes—only because, though, we know that we could easily produce a fuller story featuring psychological states.

That shows, firstly, that rationalising explanations in terms of mental states are compatible with reason-explanations in terms of facts. Secondly, that causalism about reason-explanations is compatible with externalism about reasons. And it shows, thirdly, that rationalising explanations and reason-explanations in terms of facts are not distinct at all. Rather, the latter are *elliptical*—they are short and condensed versions of the former.[25]

In conclusion we can say, firstly, that none of the views and arguments discussed in this and the previous two sections constitutes or supports a convincing objection to causalism. Secondly, we have no reason to deny that the two kinds of causalism stand and fall together. Thirdly, each of the three discussed positions on the nature of reasons is compatible with causalism about reason-explanations. And that shows, fourthly, that there is no need to argue for—or commit ourselves to—one of the discussed views on the nature of reasons.[26] I shall, therefore, in the following abstract from the distinctions made in this section. All claims about reasons and reason-explanations may be taken to be, literally, about reasons or they may be interpreted as being about reason-*states*.

---

[24] That claim concerns reason-explanations that are given in retrospect and from a third-person perspective—from an explainer's points of view. It does not concern the agent's point view at the time of deliberation or action. Further, the claim that explanations in terms of facts are short and condensed versions does not entail that explanations in terms of reason-states are primary or more fundamental.

[25] What about cases in which the agent did not have *good* reason to perform an action that can be rationalised in the light of the agent's mental states? For instance, assume that Sue's *A*-ing can be rationalised by reference to her believing that *p*, and that is not the case that *p*. The reason-state does not represent a reason—it does not represent a fact at all. Such cases are not a problem for causalism. Rather, they are a problem for the view that reason-explanations explain in terms of facts. Causalism does not require that every reason-state refers to a reason-fact. An explanation of Sue's *A*-ing in terms of her false belief is a reason-explanation, because her *A*-ing appears as rational and intelligible in the light of her believing that *p*. According to causalism, a reason-explanation may be a *mere* rationalising explanation—not a *proper* reason-explanation, if you like. The important point is that, *if* an agent was acting for good reasons, then there is no *proper* reason-explanation in terms of facts, which is distinct from a causal explanation in terms of reason-states, because the reason-states represent the reasons and because the former is a condensed version of the latter.

[26] A view that has not been discussed is based on a rejection of the dichotomy between internalism and externalism about reason. On that view, it is mistake to assume that reasons are *either* mental entities (attitudes or their content) *or* facts. Reasons, rather, should be construed as *relations* holding between facts, actions and the agent's mental attitudes (compare, for instance, Skorupski, 1999, essay 2). Construed in this way, the question whether reasons are mental states or facts does not arise. This view, it seems clear, is compatible with causalism.

The Case for Causalism

I will now argue *for* causalism about reason-explanations. That case for causalism comes in three parts. Firstly, I will outline and discuss the standard argument for causalism, which is due to Donald Davidson. Then I will turn to—and argue against—non-causalist challenges and alternatives to the causal theory. And in a third part I will present another argument for causalism that relies partly on the results from the previous chapter, and that will not be complete until the end of the next chapter. That argument says, roughly, that we should prefer causalism to non-causalism for the following two reasons. Firstly, I will show that, in conjunction with the standard-causal model of agency, causalism provides an *integrated* account of agency by locating agency in the causal order of events. And secondly, I will argue that non-causalism fails to provide a satisfying alternative account of agency.

Davidson's Challenge

Davidson asked what the 'mysterious connection' between reasons and actions consists in and how the 'explanatory force' of the 'because' in reason-explanations is to be understood.[27] Following Davidson, causalists maintain that only a causal theory can provide informative and non-circular answers to those questions. Non-causalists have criticised the causal theory and they have put forward alternative non-causal accounts.

Let us begin by considering one common way of providing a reason-explanation of an action. Suppose that Sam opens the window. Someone else asks him why he is doing that. Sam says that he opens the window in order to let in some fresh air; that is the reason why he opens it. That kind of reason-explanation has the following form.

(RE) *S A*-ed in order to *B*.

What non-causalists will point out is the *teleological* character of RE. We explain why *S A*-ed by pointing out that is was S's *goal* or *end* to do or to bring about *B*. In other words, *S A*-ed for the *purpose* of doing or bringing about *B*. They will say that the teleological character of reason-explanation stems from the goal-directedness and purposefulness of rational action, and that the burden of argument lies, therefore, with

[27] See Davidson, 1980, especially p. 9 and p. 11.

the causalists who claim that the explanatory force of a reason-explanation has to be understood in terms of causation.[28]

Causalists, however, can suggest another way of construing the reason-explanation. In answering the why-question raised above, we may as well say that Sam opened the window because he wanted to let in some fresh air. That is a perfectly natural and equally good way of giving a reason-explanation. We obtain the following form.

(REC) *S A*-ed because he wanted to *B*.

And, without committing themselves to the view that all actions issue from desires, causalists can reformulate REC in the following way.

(REC*) *S A*-ed because he desired, intended or had some pro-attitude to *B*.

What is the relation between RE and REC? What are the truth-conditions for statements of each form? Can statements of one form said to be true in virtue of the truth of the corresponding statement?

But let us begin with the question in virtue of what statements of the form REC* are true—assuming that the truth-conditions for REC* are the same as for REC, since the former is just a reformulation of the latter. Given that *S* performed an action of type *A*, and given that *S* had the attributed pro-attitude, what makes it true that *S A*-ed *for*—or *because of*—the attributed reasons? That, I take it, is one way in which Davidson's question for the explanatory force of the 'because' can be understood: to ask for the explanatory force of the 'because' in a reason-explanation is to ask in virtue of what the explanation is true.

According to the causal theory, the explanatory force of the 'because' is, partly at least, causal. That is, statements of the form REC* are true only if the mentioned pro-attitude caused the explained action. Non-causalists will respond by pointing out that the explanatory force of 'because' is not always causal in kind; there are all sorts of

[28] One may object that a teleological theory is not necessarily a non-causal theory, because purposeful activity can be understood in terms of *final* causation. But that point is merely terminological, and it presupposes the traditional—Aristotelian—terminology that distinguishes between final and efficient causation. In contemporary philosophy, however, causation usually means *efficient* causation. To say that reasons are causes is, then, to say that reasons are efficient causes. And given that all causation is efficient causation, to deny that reasons are efficient causes is to deny that they are causes—without qualification. Nothing, though, depends on the assumption that all causation is efficient causation, as all the claims and arguments could be reformulated in terms of final and efficient causation.

non-causal explanations. The question, then, is how to decide whether reason-explanations are causal or non-causal in kind.[29]

One way to settle the disagreement would be to show that reason-explanations cannot possibly be causal. An alternative and more modest strategy pursued by non-causalist is the following. First, problems that trouble any causal theory are highlighted.[30] Then, a non-causal alternative account of the explanatory force of reason-explanations is presented. And then it is argued that we have good reason to prefer the non-causalist alternative, since it avoids the problems of the causal account.[31] The first strategy was the preferred one among non-causalists before Davidson's 'Actions, Reasons, and Causes.' In that article, Davidson compellingly rejected the most influential arguments to the conclusion that reason-explanations *cannot* be causal explanations.[32] After that, most non-causalists focused on the second strategy—pointing out the following.

The causal theory provides an account of the explanatory force of reason-explanations and of the connection between reasons and actions. Without further argument, however, that constitutes a case for causalism *only if* there is no alternative account available. But that is not really an argument *for* causalism—what favours causalism here is merely the lack of an alternative. Hence, the case for causalism collapses, if it can be shown that there is a viable non-causal theory of reason-explanations. I will consider and reject three non-causalist alternatives. Before that, though, let us have a closer look at the challenge that Davidson identified for any theory of reason-explanation.

Two Aspects of Reason-Explanations

Assume that an agent performs an action for which she *has* reasons in the sense that she has intentional attitudes, which allow us to rationalise the performance of the action—the performance of the action appears as intelligible in the light of those

[29] Compare Wilson, 1989, who argues along those lines; pp. 175-183.
[30] The most troublesome problem for causal theories is probably the problem of deviant causal chains. I will turn to that problem in chapter 4, pp. 175.
[31] Compare, for instance, Sehon, 2000, who argues that non-causalism is to be preferred because it is not troubled by the problem of deviant causal chains.
[32] See Davidson, 1980, especially pp. 12-19. The most influential argument rejected by Davidson is the so-called *logical connection* argument, which was supposed to show that the logical or conceptual connections between reasons and actions exclude the possibility of them being related as cause and effect. See also Goldman, 1970, pp. 109-116 and Audi, 1993a, chapter 4.

attitudes. Davidson pointed out that to cite such attitudes does not guarantee that a successful reason-explanation is being given. He noticed that

> […] 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 [may] 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.[33]

Having a reason in favour of an action has to be distinguished from acting *for* or *because of* that reason. In other words, for a reason to be truly explanatory, it is not sufficient that it rationalises the action. It must also be case that the reason *motivated* or *moved* the agent to perform the action. It is common to make that point by using examples in which an agent had two reasons in favour of performing one and the same action. Consider an example involving an unintended outcome. Let's say that Sam opened the window, because he wanted to let a bee out of the room. The attempt failed; the bee was still around and Sam closed the window. Later on, Sam opened the window again in order to let in some fresh air, and, guess what, the bee flew out. Assuming that Sam still had the desire to let the bee out, he had reasons for the action, which are not the reasons he acted for.[34] In such cases we can ask in virtue of what only one of the two reasons the agent *had* for the action is the reason the agent acted *for*. The agent had two mental states that rationalise the performance of the action, but he acted only because of one of them. Davidson asked what the explanatory force of reasons and reason-explanations consists in. It is clear now that we have to distinguish between two aspects of that explanatory force.

Firstly, true reason-explanations explain actions in the sense that they rationalise their performance. Generally, it is rational or intelligible to perform an action only if there is something to be said for doing it—only if there is something that favours doing it. Only if, in other words, there is reason to do it. However, rationalising explanations can also be given even if there is, objectively, no reason to perform the action, because they rationalise subjectively—in the light of the agent's reasons. They

---

[33] Davidson, 1980, p. 9.

[34] Different kinds of examples could be brought forward. Consider, for instance, a case of weakness of the will. Susan had good reasons to vote for K. in the presidential elections—she believed that K. is the better candidate and she wanted to go the elections. Further, a friend of hers promised to take her out for dinner, in case she voted for K. Assume that Susan did vote K., but that she would not have gone to the elections, had she not been promised the dinner. Sue wanted go to the election and believed that K. is the candidate to vote for, but she voted *because* she fancied a nice dinner with her friend.

explain in the light of what the agent took to be a reason or, simply, in the light of what the agent wanted, believed and intended.[35] That aspect—call it the *rationalising* aspect—concerns explanatory relations between *types* of actions and reasons, because it is having reasons of a certain type that rationalises the performance of a certain type of action.

Secondly, reason-explanations explain actions in the sense that they explain why the reason motivated the agent—why the agent's being aware of certain things or being in a certain mental state motivated the agent to perform the action. We can call that aspect the *motivational* or *metaphysical* aspect, as it concerns the relation between tokens, rather than types. It concerns the connection between the performance of a particular action and the agent's being in a particular reason-state.[36]

Those two aspects can be regarded as aspects of reasons or reason-states. If an agent acts for good reasons, the corresponding reason-states rationalise and motivate the agent. If an agent acts for merely rationalising reasons, but not for good reasons, then the reason-states rationalise and motivate the agent. The two aspects, however, can also be regarded as two aspects of reason-explanations and of the explanatory force of the 'because'. A reason-explanation is true only if it rationalises the performance of the action, and only if it refers to reason-states that motivated the agent to perform the rationalised action. In other words, it must be true that the agent did the action *because of* the reasons in both the rationalising and the motivational sense of *because*.

Davidson pointed out that the explanatory force of reason-explanations is twofold in the sense explained. A theory of reason-explanation must explain what the difference between having a reason and acting for it consists in by providing an

[35] In other words, it is not required that the agent has *good* reason to perform the action. Performing a certain action may be not the best, the wrong, an imprudent, or, simply, a stupid thing to do. But we may still be able to rationalise its performance in the light of what the agent believes, desires, intends, and so on. The kind of rationality could also be called *internal* or *relational* as it concerns the relations between the contents of mental attitudes and types of actions, rather than the correctness of the agent's beliefs and judgements or the rationality of desires and intentions themselves.

[36] It is very common to introduce a distinction of that sort. The most common distinction is probably the one between *normative* and *motivating* reasons. See, for instance, Smith, 1994, and for a summary of the history of that distinction see Dancy, 2000. I find that distinction slightly misleading in two ways. Firstly, no reason is only normative or only motivating. In order to be explanatory, a reason must both rationalise the action and it must have motivated the agent. It is therefore better to talk about *aspects* of one and the same reason. Further, not all rationalising explanations of actions refer to normative reasons, but to what the agent took to be a reason for the action, or to reason-states in the light of which the performance appears as intelligible. It is therefore better to contrast the motivating aspect with the rationalising aspect, rather than a normative aspect.

account of the 'mysterious connection' between reasons and action. The challenge, in other words, is to account for both the rationalising and the metaphysical aspect of reason-explanations. In the following, I will focus on the metaphysical or motivational aspect. Causalism about reason-explanations—in conjunction with the standard-causal model of agency—construes that aspect in causal terms. An agent acted *for* a certain reason only if the associated reason-state caused the performance of the action, and a reason-explanation of that action is true only if that reason-state causally explains the action. Now I will turn to three non-causal alternative theories, and I will argue, partly in connection with the results from the first chapter, that all of them fail to account for the metaphysical aspect of reason-explanations.

A First Alternative

Consider again the three suggested forms of reason-explanation:

(RE) *S A*-ed in order to *B*.

(REC) *S A*-ed because she wanted to *B*.

(REC*) *S A*-ed because she desired, intended or had some pro-attitude to *B*.

According to a first non-causal alternative, statements of the form REC* (or REC) are true, only if the corresponding statement of the form RE is true; it is true that *S A*-ed because she wanted to *B*, only if it is true that *S A*-ed in order to *B*.

This proposal, it seems, provides a straightforward way to meet Davidson's challenge. It is true that the agent acted *for* the reasons cited in statements of the form REC*, only if the corresponding teleological explanation of the form RE is true—only if it is true that the agent had the goal of *B*-ing by *A*-ing. That is, if the agent had the mental attitudes mentioned in REC*, and if the corresponding teleological statement RE is false, then the agent merely *had* the reasons, but did not act *for* them.[37]

[37] George Wilson discusses and rejects the following response to this suggestion. Some causalists argue that statements of the form RE entail, analytically, statements of the form REC*. Moreover, our grasp of the former presupposes a grasp of the latter. If that is correct, the non-causalist alternative is profoundly mistaken, since it gets the direction of justification wrong. It is RE that needs to be grounded in REC*, rather than the other way round. Against that, Wilson argues that the non-causalist can hold that RE 'guarantees' the truth of REC* without being committed to the claim that the meaning of RE is given by—or has to be explained by—the meaning of REC*. Explanations in terms of desires, beliefs and intentions *are* available to the non-causalists, but such explanations are, according to Wilson, just redescriptions of the agent's goal or purpose of *A*-ing in order to *B* (see Wilson, 1989, chapter 7). I can agree with Wilson that statements of the form RE can usually be reformulated using

Does that constitute an alternative account of the metaphysical aspect of reason-explanations? Two points must be established. Firstly, the theory must account for the metaphysical connection between reasons and action—it must explain why a reason-explanation explains the performance of a particular action by reference to an agent's having a particular reason. Secondly, the non-causalist must explain why that alternative does not merely defer the problem. The causal theory grounds the truth of statements of the form REC* in causal connections. The alternative grounds it in statements of the form RE. But in virtue of what is a statement of the form RE true? Should we accept the fact that the agent pursued some goal or end as a bare teleological fact?

According to the causal approach, actions are events with the right kind of causal history. The performance of an action is, metaphysically speaking, nothing but the occurrence of an event. The only way to explain the occurrence of a particular event, it seems, is to provide a causal explanation. Given that, it is difficult to see how a non-causal alternative could possibly account for the metaphysical aspect of reason-explanations. However, construing the issue in that way, non-causalists will object, is to beg the question. The causalist, it seems, merely affirms the causal approach, assuming that actions are events and that only causal explanations can explain the occurrence of events.

Nevertheless, let us ask why causal explanations explain the occurrence of events. Answering that question, one may refer to causal laws. The occurrence of an event-token can be explained by pointing to, firstly, the occurrence of another event and, secondly, to a law according to which tokens of the latter type are followed by tokens of the former type. To require reference to laws, however, is also question-begging from the non-causalist's point of view. Another way of answering the question is to say that causal explanations explain the occurrence of events, because they support counterfactuals of the right sort; in our case, counterfactuals of the following form.

(CF) Given relevantly similar circumstances, had *S* not desired or intended to *B*, then *S* would not have *A*-ed.

---

statements of the form REC*—and *vice versa*. However, I do not think that dealing with this particular issue is a promising way of assessing the disagreement between causalist and non-causalist positions.

The suggestion here is to understand the 'mysterious connection' between reasons and causes as counterfactual dependence. That is clearly compatible and in line with the causal theory, since causal claims entail or support counterfactuals. Further, construing it in terms of counterfactuals does not by itself rule out non-causal theories. Rather, the suggestion is that any theory that supports counterfactuals of the form CF provides an account of the connection between reasons and action just by virtue of supporting those counterfactuals.

So, a non-causal alternative could account for the metaphysical connection between actions and reasons by showing how and why non-causal explanations support counterfactuals of the right kind. George Wilson and Scott Sehon, for instance, have pursued that strategy. Wilson thinks that teleological explanations of the form RE support counterfactuals of the form CF, and that the truth of counterfactuals of the form CF is grounded in the truth of the corresponding teleological explanation in conjunction with some fundamental and 'simple facts about an agent's power to act or to refrain from acting', such as the fact that 'agents normally have it in their power not to perform an action of a type that they have no adequate reason to perform'.[38] Given, furthermore, that they are *rational* agents, they usually refrain from performing actions they have no reason for: rational agents, typically, would not have *A*-ed, had they not had reason to do so. According to Wilson, it is simply a fundamental fact about genuinely rational agents that, if the agent *S* has no goal that would be promoted by *A*-ing, then *S* would not even try to *A*. These claims are made in the light of what Wilson takes to be further fundamental facts about agents. Firstly, human agents have the ability to *guide* and *regulate* their behaviour in a manner appropriate to the content of the desires and beliefs they have. Secondly, they can *reflect* on the value of what is presented by their desires as attractive, and, thirdly, they can choose whether they act *on* a given desire or not.[39]

Sehon's account of the explanatory power of the teleological alternative is very similar to Wilson's. Sehon discusses the role of mental states in the explanation of action and argues, partly with Wilson, that the teleological alternative can incorporate reference to desires and beliefs.[40] Sehon, however, considers different counterfactuals;

[38] Wilson, 1989, p. 198.
[39] Ibid., pp. 184-185.
[40] See Sehon, 1994 and 2000.

namely, counterfactuals that serve to characterise, as Sehon says, the explanatory power of the teleological connective *in order to*. They are of the following form.

(CF*) Had *S* not directed her behaviour towards *B*, *S* would not have *A*-ed.

But Sehon's explanation of why counterfactuals of that form hold is, basically, the same as given by Wilson in support of CF. Sehon argues as follows.

> Given (i) that it is within *S*'s power to refrain from *A*-ing, and (ii) that *A*-ing would serve no purpose toward which *S* is directing her behaviour, *S* will typically not *A*. That is a fundamental fact about the behaviour of teleologically explicable agents: so long as it is within their power to refrain, agents will typically not do things that serve no purpose of theirs.[41]

One may wonder why, exactly, such observations concerning some 'fundamental facts about agency' in conjunction with teleological statements of the form RE support the relevant counterfactuals. Why is it that the agent would not have done the action, given the antecedent? The offered answer, as I understand it, is just that *rational* agents will, usually, refrain from performing an action for which they have no reason—given that they have it in their power to refrain. And since it is, usually, within their powers to refrain, a rational agents, who *A*-ed for the reason *R*, would not have *A*-ed, if they had not had *R*.

I shall grant that the teleological alternative supports the right counterfactuals. My objection is that the proposal merely defers the problem. The proposal says that statements of the form REC* are grounded in statements of the form RE, and that the latter are grounded in the mentioned facts about rational agents; in particular, the fact that rational agents have the power to act and refrain from acting in the light of reasons. But that power is itself in need of explanation. My objection to the teleological alternative is, then, the following.

In the previous chapter I distinguished between four different positions in the metaphysics of agency: the reductive and the non-reductive approach, volitionism and pluralism. I rejected both the non-reductive approach and volitionism. In the second part of this chapter I will argue against pluralism and I will explain why the power to choose and act for reasons is in need of explanation. Finally, I showed in the first part of this chapter that the reductive approach stands and falls with causalism about reason-explanations. Given all that, we can say that there is only one viable account

[41] Compare Sehon, 1994, p. 66.

of the powers of rational agents available; namely, the account provided by the reductive standard-causal model of agency, which is, apparently, incompatible with non-causalism. That leaves non-causalism without an account of the metaphysics of agency; in particular, without an account of an agent's power to choose and act in the light of reasons.[42] We should, then, reject non-causalism about reason-explanations because it fails to account for the metaphysical aspect of rational action and reason-explanations.[43]

Wilson responds to an objection that is, in spirit, similar to the one just outlined. The causalist, Wilson says, may insist that the performance of an action has to be understood as the occurrence of an event, which, in turn, has to be explained in terms of causation. The causalist, as Wilson thinks, is in effect

> […] maintaining that the power to act in order to promote a valued objective is nothing more than the *potentiality* that the agent's having of the relevant reason will be an efficient cause […].[44]

Against that, Wilson makes the point that I indicated above. The causalist, it seems, is begging the question, since she merely reiterates, as Wilson says, the causal approach, according to which the powers of rational agents have to be understood in terms of event-causation. There is no further argument *for* the causal approach. Hence, all the non-causalist needs to do is to provide an alternative account of the powers of rational agents.

Is there a viable alternative account that is compatible with non-causalism? Wilson does not develop an alternative account. However, he offers the *idea* for an alternative theory, and he thinks that is sufficient to rebut the causalist's contention that the powers of agents must be understood in event-causal terms.[45] Wilson's idea

---

[42] Both Wilson and Sehon say that the mentioned facts about rational agents are fundamental facts, but they do not clarify what they mean by 'fundamental'. One might think that means, firstly, that the mentioned facts cannot be explained (at least not reductively), which does not matter, since, secondly, the mentioned facts are not in need of explanation. Whatever Wilson and Sehon mean by 'fundamental', they merely maintain that the mentioned facts are fundamental—that is, they do not argue that, or explain why, we should accept them as fundamental. I will *argue* further below that the efficacy of reasons and an agent's power to act on them is in need of explanation (see especially pp. 107), and the reductive account of that power presented in the first chapter will be further developed in the following two chapters.

[43] I will return to this argument at the end of this chapter.

[44] Wilson, 1989, p. 199.

[45] Wilson's aim is to offer a 'prima facie alternative' to the 'causalist views about the relation of reasons and the powers of agency' (p. 199). That suggests that Wilson does not regard the facts about agents and their powers as fundamental, in the sense that they are not in need of explanation (compare note 42).

builds on Leibniz's doctrine that reasons 'incline without necessitation'. If the reasons for which an agent acted merely inclined the agent to act in a certain way, then, as Wilson says, their 'motivating force […] has not *necessitated* the act, i.e., has not been among its efficient causes'.[46]

Wilson, however, is mistaken in several respects. Firstly, the causalist does not merely reiterate the original position. We have to distinguish between causal theories of reason-explanation and causal theories of agency. It is true that the causal theory of reason-explanation presupposes the standard-causal model of agency. But neither does the causalist simply assume that reason-explanation is a species of causal explanation, nor is it simply assumed that the standard-causal model is true. In the previous chapter, for instance, I did not assume that the powers of rational agents can only be understood by reducing them to causal relations between agent-involving states and event. Rather, I *argued* against the non-reductive model and against volitionism—and I will argue against pluralism in due course.

Secondly, it is not obvious at all that the idea that reasons incline, without necessitation, is incompatible with the causal approach. One would have to show that the notion of inclination cannot be understood in dispositional terms, or in terms of probabilistic causation—something Wilson has not even attempted to argue for.

Thirdly, and most importantly, it is difficult to see why Leibniz's claim constitutes by itself an alternative account of an agent's power to choose and act for reasons. Suppose that an agent *S* has reasons in favour of *A*-ing and reasons in favour of *B*-ing, and that both reasons merely incline *S* to choose and act. Suppose, then, that *S A*-s. How do we explain that? Do we seek to explain *S*'s choice and act by pointing out that the reasons in favour of *A*-ing were *better* or *stronger* than the ones in favour of *B*-ing, or do we say that *S A*-ed because it was within *S*'s power to choose and act in the light of reasons? What Wilson and Sehon suggest is, clearly, the latter. They think that an exercise of the power to act for reasons cannot be reduced to relations that hold between the agent's reasons and actions. Because of that, though, the qualification that the relation is one of inclination rather than necessitation does not make a relevant difference. Wilson and Sehon do not suggest that reasons render the performance of certain actions merely probable. They think, rather, that, since reasons merely incline, it is within the power and up to the agent to choose to act on them. But

[46] Compare ibid.

that is just to assume that rational agents have the ability to choose and act for reasons, rather than to provide an account of it. Far from providing the idea for an alternative theory, Leibniz's claim merely highlights what has to be explained by a theory of rational agency.

A Second Alternative

Let us now turn to another non-causal proposal. Carl Ginet has suggested another non-causalist account of reason-explanations that is teleological in character. But it differs in some respects significantly from the proposal discussed in the previous section. Consider again our reason-explanation.

(RE) *S A*-ed in order to *B*.

Instead of referring to RE as grounding the truth of other statements, Ginet offers the following non-causal—or 'anomic', as Ginet says—condition for its truth.

(NC) Concurrently with her *A*-ing *S* intended of that *A*-ing that by it (and in virtue of its being an *A*-ing) she would *B* (or would contribute to her *B*-ing).[47]

According to Ginet, the truth of NC—'besides the occurrence of the explained action'—is sufficient for the truth of RE. When these conditions are satisfied, then it was *ipso facto*, as Ginet says, *S*'s *purpose* that by *A*-ing she would *B*, which is, in turn, just to say that *S A*-ed in order to *B*. Ginet, obviously, assumes an analytic relationship between the notions of acting with a purpose and acting with an intention—an assumption that I shall not contest. Does Ginet's proposal do better than the first teleological alternative? Let us first ask whether the proposal can account for the difference between *having* a reason in favour of an action and acting *for* it.

In a first response I am tempted to say that it is *just obvious* that Ginet's proposal fails in that respect. What Davidson's challenge highlights is that an agent's reasons may merely rationalise the performance of an action—they may merely accompany its performance without motivating it. But all that NC requires is that the intention occurs *concurrently* with the action—it leaves open the possibility that their concurrent occurrence is a mere coincidence.

[47] Compare Ginet, 1990, p. 138 and 2001, p. 388.

Alfred Mele makes a related point using the following example. Suppose an agent had concurrently with *A*-ing two intentions, I_1 and I_2, that satisfy a condition of the form NC—say that according to the content of I_1, *A*-ing contributes to *B*-ing, and according to I_2, it contributes to *C*-ing. Suppose further that I_1—or its neural realisation—was causally relevant to the occurrence of the bodily movement that was or constituted the particular *A*-ing. Mele asks to which of the two intentions we would refer in order to explain the agent's *A*-ing. We would, clearly, refer to I_1, rather than I_2. That shows, according to Mele, that 'the *mere presence* in the agent of an intention' is not sufficient for that intention to be explanatory of the performance of an action.[48]

In a response, Ginet rejects Mele's example as question-begging. Asking which intention is causally explanatory, it just assumes that intentions—or their neural realisations—usually play a causal role in the performance of actions. Further, Ginet makes the epistemological point that we are, usually, ignorant about the neural processes occurring in our brain. But we are, usually, not ignorant about the truth conditions of reason-explanations, which are formulated in common-sense psychological and teleological terms.[49]

The second point is hardly convincing, though. Firstly, we are—or can be—justified in believing ordinary causal statements, such as 'the earthquake caused the catastrophe' or 'the stone caused the breaking of the window', even though we are ignorant about the underlying physical mechanisms. Secondly, Ginet assumes, apparently, that the neural realisations of intentions, rather than the intentions *qua* intentions, cause and causally explain behaviour. Causalism, however, is not committed to that claim.[50]

My response to Ginet's first point—the objection that the causalist is begging the question—is basically the same as above. Causalists do not simply presuppose the causal framework. Rather, they insist that the counterfactual dependence between reasons and causes is in need of explanation. Rejecting reference to causal connections, non-causalists have to provide an alternative account of the metaphysical connection between reasons and action that explains why the relevant counterfactuals

[48] See Mele, 1992, p. 253.
[49] See Ginet, 2001, pp. 389-390.
[50] I will say more on that in chapter 3, especially pp. 117, and in chapter 4, especially pp. 183.

are true. Ginet's proposal falls short of that. According to his view, the truth of statements of the form RE requires the truth of statements of the form NC together with *the occurrence of the explained action*. Ginet, naturally, presupposes that the action has in fact been performed. But more is needed. We want to know why the particular performance depends counterfactually on the agent having the particular intention, for instance. The fact that, *on that particular occasion*, the agent, concurrently with performing the action, intended of that action that by doing it she would pursue a certain goal, does not explain the dependence between that intention and that action. Ginet's proposal fails, therefore, to account for the 'mysterious connection' between reasons and actions—it fails to account for the metaphysical or motivational aspect of reason-explanation.

A Third Alternative

What causalists and non-causalists can agree on is that, when we explain the actions of human agents, we describe their behaviour in intentional terms and attribute intentional states to the agent, in the light of which their behaviour appears as intelligible. Seeking a rationalising explanation, we try to *understand* and *make sense* of their actions. Some philosophers think that this practice of interpretation and attribution is indeterminate, open and holistic in nature.[51] And some non-causalists think that we can distinguish between having a reason and acting *for* it by providing a sufficiently detailed and comprehensive narrative—without presupposing or appealing to causation by reasons. Frederick Stoutland, for instance, says that

> [to] determine the reasons which do explain [the agent's] behaviour, when this is not clear, we first have to determine how to describe it. This may be a complex process: we may have to ask him some questions, we may have to check out what else he did before […]; we may have to see how he behaved afterwards […].[52]

Stoutland, however, does not make explicit how such a narrative is structured and to what facts it must appeal in order to explain an action. In another article Stoutland says that the non-causal theory 'gives no explanation of why the agent's behaviour *occurs* or *comes about*'. Rather, it 'gives the attitudinal conditions in terms of which

[51] See, most prominently, Davidson, 1980, essay 11 and 12. Compare also Child, 1994.
[52] Stoutland, 1986, p. 48.

to derive the *understanding* of the agent's behaviour *as* the act that he performed'.[53] That, it seems, is just another description of what we do when we rationalise the performance of an action in the light of some of the agent's reason-states. But that means that Stoutland does not offer an alternative account at all. Rather, he merely denies that the explanatory force of reason-states must be grounded in the fact that they played a causal role in the causation of the action.

In a recent book, G. F. Schueler agrees with Stoutland on the point that a non-causal theory can account for the difference between *having* a reason and acting *for* it by virtue of providing a richer narrative. Schueler, however, makes clear that such a narrative refers to some of the agent's mental states and events *and* to certain features of the agent's character—possibly also to the agent's history of socialisation and personal development. Schueler thinks that non-causalism can provide an alternative by analogy with a solution to a 'puzzle' concerning free rational choice. He cites Thomas Nagel, who describes that puzzle as follows.

> When someone makes an autonomous choice such as whether to accept a job, and there are reasons on both sides of the issue, we are supposed to be able to explain what he did by pointing to his reasons for accepting it. But we could equally have explained his refusing his job, if he had refused, by referring to the reasons on the other side […]. Intentional explanation […] can explain either choice, […] but for that reason it cannot explain why the person accepted the job for the reasons in favour instead of refusing it for the reasons against.[54]

Without getting into detail, let us agree that in such cases we cannot explain why the agent does one thing *rather than* another in terms of the agent's reasons. According to Schueler, though, when we explain actions in psychological terms, we are not restricted to giving the agent's reasons. We might be able to explain a choice or action by referring to the agent's character or personality. By pointing out what 'kind of person' the agent is, as Schueler thinks, we may be able to understand why one action was chosen *rather than* another. Reference to the agent's personality can tell us why the agent *took* something *as* a reason for action. Reference to character traits can therefore be part of reason-explanation, because it can help us to understand why the agent choose and acted *on* one set of reasons *rather than* another.[55]

---

[53] Stoutland, 1976, p. 302, my emphasis.
[54] Nagel, 1990, pp. 115-116. Cited in Schueler, p. 50 and p. 84.
[55] Schueler, 2003. For a summary of the view that reference to character traits can explain why the agent takes something to be a reason see p. 81.

Schueler's suggestion is to make use of the same explanatory resources in order to meet Davidson's challenge. The thought is, it seems, that we can explain the difference between *having* a reason and acting *for* it in the same way as we can explain why an agent acted for one reason *rather than* another.[56] Both questions can be answered by supplementing the reason-explanation with considerations concerning the agent's character, which tell us why an agent *took* something to be a reason—which is to say that Davidson's challenge can be met without appeal to causation or causal explanation.[57]

I think, though, that Schueler misses the point. The first thing to note is that in the case of rational free choice the agent has different reasons in favour of *different* courses of action, whereas the question for which reason an agent has acted arises also when the agent has different reasons in favour of *one* course of action. Far more importantly, though, Davidson's challenge does not concern the question why an agent acted for one reason rather than another. Rather, it asks what acting for a reason consists in: given that the agent acted for a reason, what makes it true that the agent acted for that reason? What is the metaphysical or motivational connection between that reason and the performance of the action in virtue of which it is true that the agent acted because of having that reason? Examples in which an agent has more than one reason in favour of *one* action can be used to illustrate the problem. But the challenge concerns acting for reasons in general—it concerns just as well cases in which there is only *one* reason that favours the performance of *one* action.[58] Schueler thinks that the challenge is about 'gaps' in intentional explanations of action. But that is a misunderstanding. The challenge is neither to fill in explanatory gaps, nor to explain how we can provide richer explanations of actions. Rather, *given that* reference to a certain mental attitude is explanatory, the challenge is explain why

[56] For the analogy between 'Davidson's challenge' and the problem of rational free choice see ibid., pp. 84-87.

[57] Note that it is far from obvious that explanations in terms of character traits, construed as dispositional properties, cannot contribute to causal explanations (reference to an object's fragility, for instance, can contribute to a causal explanation of that object's breaking under certain conditions).

[58] Compare Child, 1994, who points out that the problem does not only arise in cases in which we have to judge *which* reason the agent for, but it arises for every case in which an agent acted *for* a reason (especially p. 96).

referring to it is explanatory—the challenge is to provide an account of the rationalising *and* the motivational aspect of reason-explanations.[59]

Agreement between Causalists and Non-Causalist?

In a more recent article Stoutland acknowledges that point. Reference to causation is not supposed to help us to find out which reasons an agent acts for, nor is it supposed to explain why the agent acted for one reason rather than another. It is not even meant to explain why the agent acted for a reason, in case there was only one reason favouring one action. Rather, causalism says that a causal relation is a necessary condition for acting for a reason. It is a condition, as Stoutland says, that 'must be met if the claim [that the agent acted for a reason] is to be true, not an explanation of why it is true'.[60]

Curiously, Stoutland does *not* argue against that claim, and he does not attempt to provide an alternative. Rather, Stoutland considers Davidson's position in comparison to *intentionalism*, which is the version of non-causalism that he holds, and he argues that there is hardly any disagreement between the two views—at least far less disagreement than it is usually thought. In particular, intentionalists need not deny that there is a causal connection between *some* mental events and actions that are done for reasons. This is because intentionalism denies only that there are causal connections between actions and the mental attitudes that rationalise their performance. These claims require clarification.

Stoutland agrees with Davidson that actions are events, and that the occurrence of an event has to be—or is best—explained by citing its cause. He agrees, further, that causal relations hold only between events. That is, to refer to an event's cause is to refer to another event. Given that, it is only plausible to assume that actions are caused by events, and it is plausible to assume that those events are *mental* events,

---

[59] One may think that the difference between having a reason and acting for it can be explained in terms of the agent's *treating* the corresponding consideration as a reason. But that is not very promising. Reference to a particular treating of a consideration as a reason raises more questions than it answers, and it merely defers the challenge. If we say that the agent performed a particular action *because* she treated a certain consideration as a reason, the question is what the explanatory force of *that* 'because' consists in. Further, the proposal raises the question whether treating as a reason is itself an *action*. If so, is that mental action itself done *for* reasons? Finally, the present proposal is, basically, the same as the neo-Kantian proposal that acting for reasons consists in the agent's endorsement of a certain motive as a reason. I argued in the first chapter that this view does not give us a real alternative, because the agent's power to act in accordance with the endorsement is itself in need of explanation.

[60] Stoutland, 1998a, p. 197.

since we are concerned with *intentional* action. Further, Stoutland agrees with Davidson on the point that intentional actions are explained in terms of the agent's reasons, in the sense that the explanation attributes mental attitudes that rationalise the performance of the action—in the circumstances. According to Davidson, though, the mental attitudes that rationalise actions are dispositional *states* and not events. So, why does Davidson think that reason-explanations are causal explanations?

According to Davidson, the mental attitudes that rationalise the action can usually be 'associated with' the mental event that causes the action.[61] There are two obvious ways in which that relation can be construed. Either the event is associated with an attitude in the sense that the event is the becoming *occurrent* of a standing attitude—the manifestation of a disposition to desire to *A* in certain circumstances, for instance.[62] Or the event is associated with the agent's having of a certain attitude at a time.

Stoutland assumes that, for every action *A* and for every mental attitude *R* that rationalises *A*-ing, there is a mental event that can be associated with *R* and that causes the *A*-ing.[63] He points out, though, that it does *not* follow that the corresponding reason-explanation in terms of *R* is a *causal* explanation, since it is not obvious that the attitude is causally explanatory just because it can be *associated* with the event. Further, he thinks that there is no need for intentionalists to deny that actions are caused by mental events (which can be associated with attitudes that rationalise them). But Stoutland also thinks that there is no reason to assume that either, because he fails to see what would be gained by insisting on it.[64]

However, I think that there is good reason to endorse that claim. What is gained is an account of the metaphysical relation between reasons and actions: actions are caused by mental events, which can be associated with mental attitudes that rationalise the performance of that action. That account is certainly not unproblematic. But having a problematic account, it seems to me, is preferable to having no account at all.

---

[61] Stoutland refers to Davidson, 1980, p. 12.
[62] I think that is what Davidson had in mind when talked about the *onslaught* of mental states or attitudes (1980, p. 12). For the notion of *occurrent* mental states see, for instance, Goldman, 1970, pp. 86-89, and Mele, 2003, pp. 30-33.
[63] Stoutland is sceptical whether there is in fact for every action, which can be rationalised by attributing a mental attitude, a mental event that can be associated with that attitude. However, he does not *argue* against it, and he assumes it for the sake of the argument.
[64] Stoutland, 1998a, especially pp. 204-205.

Stoutland does not offer an alternative non-causal account of the metaphysical aspect of reason-explanation, nor does he *argue* against the need of an alternative account. That alone, I think, gives intentionalists good reason to endorse the outlined causal account. Another reason is that the causal theory provides an integrated account of rational agency, in the sense that it tells us how the causes of the agent-involving event that constitutes an action relate to the reason-states that rationalise its performance.[65] Given that, a lot is gained by endorsing the causal account of the metaphysical aspect of reason-explanation. There is, then, good reason for intentionalist to endorse that account, given the absence of non-causal alternatives.

Reasons and Causes

Causalism says that reason-states causally explain actions. That presupposes that reason-states are causes in some sense. Apart from the suggestion that reason-states can be *associated with* mental events that cause the action, the following two further options are available to causalism.[66]

Firstly, one may deny that only events are causes. Standing dispositions, states and other background conditions often play an important role in the production of an effect. Often, reference to an event is explanatory only because we know that certain other conditions obtained. These standing background conditions, so the suggestion goes, are not merely explanatory, but they play a role in the causal history of the effect. And often it is relative to interest whether we refer to an event or a standing condition as *a* cause—or *the* cause—of the effect.[67]

Secondly, one may deny that there is a clear distinction between events and states. It has been suggested to construe *both* events as states as instantiations of

[65] I will develop this point in more detail in the next section and in chapter 3.

[66] In the following, one must bear in mind the conclusion of the section on externalism about content (pp. 64). In particular, a reason-state may be a cause of an action only in the sense that the description of that state refers to a physical event part of which is the cause of the agent-involving event that constitutes the action. In other words, the cause of the agent-involving event that constitutes the action may only be *part* of the supervenience base of the reason-state.

[67] That point goes back to John Stuart Mill, 1846, at least. Compare, for instance, also Lewis, 1986, p. 162, and Dretske, 1988, who argues that 'we can divide things up as we please. The fact is that in ordinary affairs we seldom, if ever, regard the cause of an event as the *totality* of conditions relevant to the occurrence of the effect. Instead, we pick out some salient parts of this totality and designate it as the cause' (pp. 39). According to Dretske, both standing conditions (states and dispositions) and events are part of the totality of an effect's causal history; he calls the former the 'structuring causes' and the latter the 'triggering causes' of the effect (pp. 42).

properties (by a substance at a time).[68] On that view, the only difference between events and states is that states are instantiated over some extended period of time. But since events and states fall under the same ontological category, there is no obvious reason to deny that they can be causally efficacious and relevant in the same way—namely, as instantiations of properties.[69]

The Metaphysics of Reason-Explanation

I argued that Davidson's challenge has to be understood as a metaphysical challenge. It concerns the motivational efficacy of reasons and the metaphysical connection and dependence between reasons and actions. A response to the challenge must show *why* and *how* reasons influence the actions of rational agents and what the connection and dependence between them consists in. I argued that none of the non-causal proposals provides a viable alternative to the causal theory.

Non-causalists, though, may respond as follows. It is true that non-causal alternatives do not show what the 'mysterious connection' between reasons and actions consists in—they do not provide a metaphysical account of the influence or efficacy of reasons. But why is the metaphysical connection between reasons and action in need of explanation at all? And secondly, if it is in need of explanation, why must a theory of reason-explanation provide an account of the metaphysical connection between reasons and actions?

One answer to the first question has already been given in the response to the first non-causal alternative. Non-causalists do not deny that actions depend counterfactually on reasons. Why are those counterfactuals are true—what *makes*

[68] Compare Kim, 1993, pp. 33-34.

[69] According to an alternative, it is sufficient to point out that states, dispositions and other standing conditions are causally *explanatory* in the sense that they play an indispensable role in causal explanations. Usually, reference to the event that triggered the effect is explanatory only in combination with the background conditions. Often the background conditions are not mentioned in causal explanations. But that is only because we take the presence of the right conditions for granted. (Suppose one explains *that* lighting of the fire by *that* striking of the match. The fact that we do not have to mention the presence of oxygen in the explanation does not mean that it is not causally explanatory. It means only that common sense takes it for granted that this condition obtained.) Endorsing that third position, causalists need not abandon the assumption that only events are proper causes, but they must give up the claim that reason-states are causes altogether. That is why the other two options are preferable. Reason-states, it seems to me, should not merely be causally explanatory, but they should be causes, in the sense that they initiate or trigger actions. Note, though, that mixed views are possible. One may hold, for instance, that an agent's standing background beliefs and desires are merely causally explanatory, whereas formations of intentions are causes of actions.

them true? The metaphysical connection between reasons and actions is in need of explanation, because the counterfactual dependence between them is in need of explanation. In the following, I will provide another response to the first question and the second part of my response to the second question. The first part of that response goes as follows.

We can grant that a theory of reason-explanation need not itself provide an account of the metaphysics of acting for reasons. However, a theory of reason-explanation *presupposes* a metaphysical connection between reasons and actions—it presupposes that reasons are motivationally efficacious in the production or performance of actions. Explanations in terms of reasons make sense only if it is assumed that the reasons motivated the agent. The fact that reasons influence the actions of rational agents is a presupposition without which reason-explanations would be entirely unintelligible. That is not to say that we cannot make sense of an agent's behaviour without knowing what she actually had in mind when she was acting. We may try to interpret her actions by attributing mental attitudes in the light of which the agent's behaviour makes sense. But in doing so, we assume that the agent was motivated by those attitudes—we are assuming a metaphysical connection and dependence between reason-states and actions.

A theory of reason-explanation does not stand on its own feet. It is part of a theory of agency in general, and part of a theory of acting for reasons in particular. In the first chapter we saw that only some positions in the metaphysics of agency are compatible with the causal theory of reason-explanations, and that others are committed to non-causalism. So, even if a theory of reason-explanation need not itself provide an account of the metaphysics of acting for reasons, it must be compatible with such an account. And, in general, a theory of reason-explanation should not be assessed as a freestanding theory, but as part of a theory of agency.

Causal Closure and the Efficacy of Reasons

Consider now another explanation of why the efficacy of reasons and the metaphysical connection between reasons and actions is in need of explanation. Let me begin by introducing the principle of the *causal closure of the physical*. The principle says, according to Kim, the following.

> If you pick any physical event and trace out its causal ancestry […], that will never take you outside the physical domain.[70]

Non-causalists, usually, have no intention to deny or question the closure principle. And together with that principle they usually accept that bodily movements can be explained in scientific—and causal—terms.[71] The closure principle leads to very difficult problems concerning the causal role of reason-states and concerning the status of reason-explanations. Given that the causal history of every instance of behaviour can be spelt out in purely physical—or, say, neuro-physiological—terms, there is, it seems, no causal role left that reason-states could play in the production of action. And given that behaviour can be fully explained in scientific terms, reason-explanations appear to be redundant. The first problem is known as the problem of causal exclusion, and the second as the problem of explanatory exclusion.[72]

One may think, though, that the exclusion problems arise only for *causal* theories of agency and reason-explanation. The closure principle threatens to exclude the efficacy of reason-states only if their efficacy is construed as causal efficacy. Further, it seems that two explanations can exclude each other only if they explain the same phenomenon and only if they are explanations of the same kind—only if they are two causal explanations of one and the same event, for instance. Non-causalists may argue, then, that the closure principle does not pose a problem for their view for the following reasons.

Kim, for instance, argued that two causal explanations of one and the same phenomenon create an *unstable situation*—an epistemic *tension*.[73] The reason for that is that both explanations explain the occurrence of one and the same phenomenon. According to non-causalism, however, scientific and reason-explanations of behaviour are not of the *same kind*. The former explain why certain events *occurred* in causal terms, whereas the latter explain why certain actions have been *performed* in terms of the agent's reasons. The former explain an occurrence by reference to other events, and the latter rationalise the performance of an action by an agent. Given that,

---

[70] Kim, 2000, p. 40.
[71] Dupré, 2001, and Merricks, 2001, deny the causal closure of the physical. Both advocate an agent-causal theory of agency. It is not clear, though, whether they are causalists or non-causalists about reason-explanation. But that does not matter, since I rejected agent-causation on different—and independent—grounds.
[72] I will turn to the problem of causal exclusion in chapter 3. For the problem of explanatory exclusion see, for instance, Kim, 1997.
[73] Kim, 1997, p. 265 and p. 272.

it is not obvious at all that the two explanations exclude each other—or that they are somehow in competition—simply because reason-explanations do not explain occurrences at all. A closely related second reason is that, according to non-causalism, scientific and causal explanation do not explain one and the *same thing*. Reason-explanations explain the performance of *actions*, whereas scientific explain the occurrence of *bodily movements*.

For the moment, let us set aside considerations about causal and explanatory exclusion.[74] Given the causal closure of the physical, we can see why the efficacy of reasons and an agent's power to act on them is in need of explanation. In particular, it is in need of explanation how the efficacy of reasons *relates* to the causal efficacy of the causes of bodily movements.[75] Further, it seems clear that the two non-causalist responses, which have just been outlined, do not help in that respect. Reasons undoubtedly do influence our actions. We do certain things, rather than others, because of the reasons we have—because we believe, desire and intend *this* rather than *that*. And had we had, in a particular situation, different reasons, we would have acted differently. In that plain and straightforward sense, reasons are efficacious. Given, though, that the causal history of behaviour consists only of non-mental states and events, the question how and where reason-states enter into the picture is only a natural question to ask. How is their influence is to be construed? How do they relate to the causes of movements? What does an agent's power to act on them consist in?[76]

Now we can see that the response that reason-explanations are not causal in kind is beside the point. The question how the influence of reasons is to be understood arises for every theory of agency. Further, the point that reason-explanations explain actions rather than bodily movements is also of little help. Actions, it seems obvious, stand in some intimate relationship with bodily movements; overt actions are identical with or in some sense constituted or realised by bodily movements. Given that there is an intimate relationship between how we *act* and how we *move*, it is seems natural to think that reasons must influence movements insofar as they influence actions—in

[74] In chapter 3 I will return to the problem of causal exclusion and the question whether reason-explanations are reducible to non-psychological explanations.

[75] Compare also Kim, 1997, who argues that the tension between two competing explanation can be resolved by providing an explanation of how they are related to one another (p. 272).

[76] Making that point, the causalist is not begging the question. It is not assumed there that must be a *causal* connection between reasons and actions, since reasons must have an effect on the way we move and act. The causalist merely highlights that reasons do influence movements and actions. To say that this motivational influence is causal influence is a further and independent claim.

order to have an effect on the way an agent acts, reasons must have an effect on how that agent moves. Several clarifications are in place, though.

Overt Actions and Mental Actions

We distinguished between overt and mental actions. The former are actions that involve bodily movements such as raising one's arm, and the latter are acts such as making a decision or solving a puzzle in one's head. The relationship between reasons for actions and the causes of *movements* concerns, therefore, only *overt* actions. However, in the first chapter I pointed out that both kinds of actions can be construed as being identical with or constituted by agent-involving events. In the case of overt actions the relevant agent-involving events are bodily movements, and in the case of mental actions the relevant agent-involving events are mental events (mental occurrences or the agent's having of thoughts). Given that, it seems clear that all considerations concerning overt actions and bodily movements can be carried over to mental actions and the relevant agent-involving mental events. In particular, given that mental actions are identical with or constituted by mental events just as overt actions are identical with or constituted by movements, mental actions stand in the same intimate relationship with agent-involving events as overt actions.[77] In general terms, our question is, then, how the relationship between reasons for action and the relevant agent-involving events has to be construed. (Given that straightforward connection, I will restrict my considerations to overt actions and bodily movements.)

Events and Processes

In the first chapter I pointed out that we have to distinguish between component and process versions of, for instance, the standard-causal and the agent-causal model of agency. Both the standard- and the agent-causal theory identify causal processes constituted by, what I called, an effect-component and an antecedent-component. According to the standard-causal model, the antecedent-components are agent-involving reason-states, and according to the agent-causal model, the antecedent-component is the agent—the person or the self. The effect-component is on both views an agent-involving event—a bodily movement, for instance. According to the

[77] One dissimilarity is that the events that constitute mental actions are *mental* events, whereas the ones that constitute overt actions are bodily movements (that is, non-mental events). However, provided that mental events are realised by physical events, that difference is not significant.

component version of either model, the action is identical with or constituted by the effect-component. The action, for instance, is identical with or constituted by a bodily movement with the right causal history. According to the process versions, the action is identical with or constituted by the process—the antecedent's causing a bodily movement in the right way, for instance.[78] One may think that the process view, if correct, raises a difficulty for causalism. According to the causal theory, reason-states cause and causally explain *actions*. However, that claim is plainly false, if actions are constituted by the causal processes in question.

My response to that challenge is that there is no convincing reason to adopt the process view.[79] The only obvious reason for *non-causalists* to adopt the view is that it raises an apparent problem for causalism. But that, of course, is not a reason to endorse the view, as it is, in the present context, not an independent reason. Dretske, however, argued that the process view is preferable, because it captures better where action 'begins' and where it 'ends'. He says that to identify behaviour

> […] with a process avoids [problems] by making behaviour begin where it should begin (with those efferent activities that bring about movement) and end where it should end (with those external events or conditions that the behaviour requires for its occurrence). A person's moving his arm is then a piece of behaviour that begins with those internal events producing arm movements and ends with the arm movements they produce.[80]

However, as Alfred Mele as pointed out, a proponent of the component view need not deny at all that action begins with 'efferent activities that bring about movement',

---

[78] Note that there are two notions of *constitution* in play here. According to the first, an action is constituted by an *event*, according to the second, it is constituted by an event-causal *process*. I say that, according to the component view, actions are identical with *or* constituted by events, because according to a fine-grained view of events (compare note 28, pp. 19) actions cannot be identified with non-actional events. I will say more on that further below, pp. 99.

[79] Another possible response is to reject the assumption that actions are, metaphysically speaking, *either* processes *or* components. In the first chapter we compared actions to banknotes and sunburn. What they have in common is that their causal history is constitutive of their identity. Segal and Sober, for instance, suggest that it is a mistake to ask with respect to such historical phenomena whether they are processes or products. Rather, they are *both* processes and products (see Segal & Sober, 1991, p. 22). The fact that sunburn is a condition of the skin that is, essentially, caused by exposure to sunlight suggests that it is a process. However, it is also undeniable that exposure to sun causes sunburn, which suggests that it is a product. We should accept, therefore, that terms like 'sunburn' denote ambiguously; we can use them in order to refer to the process or the product. Given that, it would be foolish to *stipulate* that sunburn is a process in order to *show* that exposure to sun does not cause or causally explain sunburn. And it would be equally foolish to stipulate that action is a process in order to *show* that causalism is false. According to that view, it is correct to say that actions are processes. But it is also correct to say that reason-states cause and causally explain them, because actions can also be identified with the effect-components of the processes in question.

[80] Dretske, 1988, p. 17- 18.

because the 'efferent events' that trigger bodily movements may be part of a causal chain that is caused by the reason-states (they may be part of the causal process that is caused by agent-involving reason-states). In that case it is, strictly speaking, true that action is a causal process. But it is not a process that involves the mental antecedents of action. Rather, it is a process that is constituted by the bodily movement and, say, certain events involving the agent's nervous system. On that view, the effect-component is a process rather than an event. It is, therefore, a component rather than a process view.[81] Proponents of the component view are not committed to the claim that actions are *simple* events. Actions may well be complex processes or chains of events.[82] All the components view denies is that the mental states and events that cause and rationalise actions are *part* of the processes that constitute actions.[83]

Another possible argument for the process view says, roughly, that reason-states cannot be causal antecedents of action, because an agent's reason-states *guide* intentional actions, rather than merely causing or triggering them.[84] The thought is that reason-states cannot be causal antecedents, which precede actions, because the *performance* of the action must be guided and monitored by the agent's mental states and events.

Let us can grant that this is required with respect to *some* actions—it is rather questionable that the performance of all actions must be guided and monitored. It is difficult to see why that would favour the process view. The mental antecedents of actions are not necessarily antecedents, which literally *precede* the action in the sense that they cease to exist at the time the action begins. According to causalism, actions are, typically, caused and rationalised by desires, beliefs and intentions. It would be absurd to think that the agent ceases to have the relevant desires, beliefs and intentions as soon as she starts to perform the action, which is caused and rationalised by them. Some of those mental attitudes are dispositional states, which contribute

[81] Brand, 1984, has suggested such a view. Brand, though, thinks that the processes that constitute actions are themselves complex or structured *events* (p. 16).
[82] In particular, some non-basic actions may well be construed as causal processes involving more basic actions, rather than as simple events. For instance, Brutus' killing Caesar can be construed as a process involving Brutus' stabbing Caesar and Caesar's death.
[83] Searle, 1983, distinguishes between prior intentions and intentions-in-action and he thinks that the former cause the latter, which in turn cause bodily movements. The action is identified with the causal process between the intentions-in-action and the movement. That is also not a proper processes view, as the prior intention, which causes the action, is not construed as being part of the process that constitutes the action (see especially p. 94).
[84] Compare, for instance, Frankfurt, 1988, pp. 73-75.

causally in the performance of the action, and others are proper events, such as an agent's believing or intending something at a time. Even for reason events, there is no reason to think that the agent ceases to have the belief and intention, for instance, as soon as the action is caused. Rather, the reason-states and events may causally *sustain* the action, which would account for their guiding and monitoring its performance.[85] Or the action is construed as a chain of events, the progression of which is guided and monitored via so-called causal feedback loops that hold between those events and the agent's having the belief and intention in question.[86]

The two considerations presented do not give any independent reason to prefer the process view, and I am not aware of any other argument for it. Is there any reason to prefer the component view? Consider an agent, Sam, who had to decide whether to *A* or *B*. After considering the reasons for and against both *A*-ing and *B*-ing, Sam judged that would be better to *A*; he then formed an intention to *A* in appropriate circumstances; and, finally, he carried out that intention accordingly. According to the standard-causal theory, that process is a causal process in which the reasons cause the judgement, which causes the intention and finally a bodily movement. If we construe all actions as processes, we face difficult questions concerning the beginning and the end of those processes. In the example, Sam performed two actions: he made a decision and he performed an overt action. Where do the two actions begin and end? If the antecedents of the overt action are the reason-states on which the judgement is based, is then the decision a proper part of the overt action? Given that there is no independent reason to prefer to process view, I think we have reason to endorse the component view just in order to avoid such questions. For, on the component view, the decision is, simply, the formation of the intention, and the overt action is the bodily movement.

Actions and Events

In the previous section I argued for the component view, according to which actions are either identical with or constituted by non-actional events. Whether or not actions can be identified with non-actional events depends on how events are individuated. Let us first have a closer look at the identity view. What does it mean to say that

[85] For more on the notion of sustaining causation see, for instance, Audi, 1993b.
[86] For more on the notion of guidance by causal feedback loops see, for instance, Bishop, 1989, pp. 168-171, and Mele, 2003, pp. 55-58. See also chapter 4, p. 178, note 49.

actions *are* events? Consider, once more, the action of Sue's raising her arm and the event of Sue's arm rising. Does the event occur whenever Sue performs the action: is Sue's arm rising whenever she raises it? The answer, it seems, is clearly *yes*. The identity view does justice to that intuition. According to that view, the agent-involving *act-neutral* or *non-actional* event of Sue's arm rising is an action in virtue of being caused in the right way by some of Sue's reason-states. The particular action—the act-token—is identical with the event-token, and the fact that a token of the type *raising one's arm* has been performed entails that a token of the type *an arm's rising* has occurred. The agent-involving event-token, as John Bishop has put it, is a token of an event-type that is *intrinsic* to the action-type:

> Each type of action is a bringing about by the agent of a specific type of event or state, and this is what counts as the event- or state-type intrinsic to that action. […] Thus, events and states intrinsic to action are always open to descriptions under which they have an unproblematic place in a naturalist ontology […]. A reference to an event-type *intrinsic* to an action is not a reference to any kind of action.[87]

An identification of act-tokens with non-actional event-tokens, however, presupposes a theory of events which is not entirely unproblematic. According to that view, events are concrete particulars, which not only can be described and referred to in different ways, but have different properties. In the example, the particular has the property of being a raising of one's hand and it has the property of being a rising of an arm.

---

[87] Bishop, 1989, p. 105. Compare also Enç, 2003, who uses the more technical notion of the *result* of an action, which is taken from von Wright and McCann, in order to capture the idea that actions are constituted by events (see Enç 2003, p. 9 and pp. 85-88). Moya, 1990, argues that there are many actions, which do not have intrinsic events, such as making an offer or giving a lecture (pp. 38-40). There are two responses. Firstly, the fact that we find it difficult to describe the intrinsic event in such cases is a linguistic obstacle; it does not show that there is no such event (compare Enç, 2003, p. 87, note 87). Secondly, the mentioned actions are non-basic actions. There is, however, no reason to deny that all *basic* actions have intrinsic events. We may modify the thesis accordingly and say that all actions either have intrinsic events or are generated by actions that have intrinsic events. That is sufficient to rebut the challenge.

Further, one may think that omissions or so-called negative actions, such as not voting or not saving the person in need, raise a problem for that view, as there is no positive occurrence that is caused by the agent's reason-states. However, I think that talk about 'negative actions' is elliptical. In some cases, the action is simply the agent's deciding not to perform an action (or to do something else instead). The remaining cases can be divided in basic and non-basic cases. Basic omissions, such as not raising one's hand, are usually constituted by the agent's actual bodily movements (whereby the "movement" may be simply holding the body still). Non-basic omissions are, generally, generated by basic omissions; not voting, for instance, may be generated by not raising one's hand. There are, no doubt, many subtle and difficult questions concerning omissions that a theory of responsibility must answer, but I cannot see why omissions would raise a particular problem for the view that actions are events. For more on responsibility for omissions see, for instance, Fischer and Ravizza, 1998, chapter 5, and for more on negative actions see Mele, 2003, pp. 146-154.

According to a rival theory of events, particular events are identical if and only if they involve and are constituted by the same substance, the same property and the same time.[88] That view construes events as property instantiations; in particular, every event is the instantiation of only one property by a substance at a time.

Kim pointed out that we must distinguish between two kinds of properties that are associated with an event. There is the property that is *constitutive* of the event, and there are numerous further properties that are instantiated *by* the event.[89] Consider, for instance, Sue's raising her hand. On the property instantiation view, that action is an event, which involves Sue, the property of being a raising of one's hand, and a certain time. The property of being a raising of one's hand is constitutive of the event. But that event also instantiates other properties such as being an action that is performed on planet earth, being an action that involves Sue, and so forth.

So, what if instantiations of the property being a raising of one's hand and instantiations of being a rising of a hand constitute *distinct* events? Even if so, it is obvious and undeniable that they stand in some intimate and systematic relationship. That fact that Sue raised her hand entails that the event of her arm's rising occurred, and the action, it seems clear, was in some sense *constituted* or *realised* by that act-neutral event.[90]

Further, it should be noted that it is at least possible that two particulars are not identical *and* not entirely distinct. According to Kim, two events may be distinct—and hence not identical—without being *entirely* distinct. In an example Kim compares the event *Sebastian's stroll* with *Sebastian's leisurely strolling*. Are there two events happening at the same time, or is there only one event with different

---

[88] Compare, for instance, Kim, 1993, essay 3 and Goldman 1970, chapter 1.

[89] Kim, 1993, p. 43: 'the properties an event exemplifies must be sharply distinguished from its constitutive property (which is exemplified, not by the event, but by the constitutive substance of the event)'.

[90] Alternatively, one may deny that a rising of Sue's hand occurs when she raises her hand. That is, as far as I understand it, Goldman's view. According to Goldman, 1970, actions are instantiations of act-properties. Sue's raising her hand is the instantiation of an act-property by Sue at a time. That action, to be sure, *is* an event (because it is the instantiation of a property by a substance at a time). But it is the instantiation of an *act*-property. And when Sue performs that action, the event of her arm's rising does not take place at all. However, I find the idea that Sue's arm is not rising, when she raises it, very counterintuitive. In absence of further and independent argument, the view that I suggested in the text should be preferred, simply because it preserves the intuition that Sue's arm rises whenever she raises it. More importantly, though, Goldman's view is committed to the claim that reason-explanations and scientific explanations of behaviour explain one and the same thing under the *same description*—both explain Sue's raising her arm. That strikes me as plainly false. Scientific explanations explain events *as* event-types—they explain, for instance, bodily movements described as bodily movements (and not described as actions).

properties? That depends on whether *strolling* and *strolling leisurely* are constitutive properties. According to Kim, Sebastian's stroll and Sebastian's leisurely strolling are distinct events. But they are not *entirely* distinct, since the latter event, as Kim thinks, metaphysically *includes* the former.[91]

Pluralism and a Miraculous Coincidence

Causalism and non-causalism disagree on the question of whether the efficacy of reasons is causal in kind and whether the relationship between reasons for actions and the causes of bodily movements is in need of explanation. Non-causalists deny that this relationship is in need of explanation—they endorse what I have called *pluralism*.

Generally, pluralism is a view concerning the relationships between different levels of explanation or different domains of discourse.[92] The behaviour of human beings, for instance, can be explained in terms of reasons for action and in terms of the causes of their bodily movements. Reason-explanations belong to the level of psychological explanation, which employs mental or intentional vocabulary. Causal explanations of movements belong to the level of, say, neurology and physiology, which employ only non-mental vocabulary. Pluralism, as I understand it, says that different levels of explanations—or difference domains of discourse—are *autonomous* in the sense that the relationship between them is *not* in need of explanation. In our particular case, to assume a pluralist position is to deny that the efficacy of reasons is in need of explanation in the sense that we do not need to explain how the motivational influence of reasons relates to the causal efficacy of the neurological and physiological antecedents of bodily movements. In particular, it is the view that the explanatory of force of reason-explanations need not be vindicated by showing how the efficacy of reason-states relates to the causes of bodily movements. Rather, the attempt to understand that relationship between the domain of intentional explanation of action and the domain of causal explanation of bodily movements is dismissed as a misguided metaphysical project.[93]

---

[91] Kim, 1993, essay 3.

[92] I will explain the model of levels of explanation in more detail in chapter 3, pp. 113.

[93] For a similar characterisation of pluralism see Cussins, 1992. Few philosophers are self-proclaimed pluralists in that sense. But there are many positions, which clearly fall under pluralism. Putnam, for instance, says that he does not reject the '[thesis or question of psychophysical correlation], *but the idea that the question makes sense*. [The] very picture that is presupposed by the question is wrong, that is to say, the picture of our psychological characteristics as "internal states" that, qua internal

Pluralism, however, is a deeply unsatisfactory position. And non-causalism is an unsatisfactory position insofar as it is committed to pluralism. In the following I will explain, firstly, why pluralism is unsatisfactory, and then why non-causalism is committed to the pluralist stance.

In the preceding sections we considered in some detail different ways of construing the intimate relationship between overt actions and bodily movements (or, in general, between actions and their intrinsic events). On all the positions considered, the relationship can be described as *intimate* and *systematic*. That is, of course, no surprise. Sue's raising her arm is necessarily accompanied by the rising of Sue's arm. According to the identity view, every instantiation of the act-type *raising one's arm* is identical with an instantiation of the event-type *rising of the arm*. On the process view, every instantiation of the act-type is partly constituted by an instantiation of the event-type. According to the third position, the act-token includes the event-token, if they are distinct, but not entirely distinct particulars. And if they are entirely distinct, the act-token is dependent on and realised by the intrinsic event-token. The important point is that according to all positions the relationship is systematic in the sense that it is no coincidence that the instantiation of a certain act-type is accompanied by the instantiation of a certain non-actional event-type. In other words, the act-token is not simply accompanied by an event-token of a certain type, but it is identical with, constituted or realised by the instantiation of a certain event-type. Given, firstly, that bodily movements of rational agents can be explained in neuro-physiological terms, secondly, that reasons influence the actions of rational agents, and thirdly, that actions are identical with, constituted by, or realised by bodily movements, the relationship between the efficacy of reasons and the causal efficacy of the neuro-physiological antecedents of movements is systematic and not coincidental—and that is why the

---

states, must either be "correlated" or "uncorrelated" with what goes on inside […] our bodies' (1999, p. 132). Burge, 1993, says that '[mentalistic] explanation and mental causation do not need validation from materialist metaphysics' (p. 117). Similarly, Baker, 1993, argues that 'systematic explanatory success [in scientific and common sense psychology] stands in need of no metaphysical underpinning' (p. 94). There are close affinities between what I call pluralism and so-called *instrumentalism* in the philosophy of mind (Dennett, 1989) and *pragmatism* in the philosophy of science (Rorty, 1979). As pointed out in chapter 1, Kant's doctrine of two worlds—or two standpoints—can plausibly be interpreted as a version of pluralism. Finally, some positions in the theory of action, which are inspired by Wittgenstein's later philosophy, can be classified as forms of pluralism; examples are Anscombe, 1957; Melden, 1961; Stoutland, 1976; and von Wright 1974.

relationship between the domain of intentional explanation of actions and the domain of non-intentional explanation of bodily movements *is* in of need explanation.[94]

Pluralism does not even *fail* to provide an explanation of the relationship between the domain of reason-explanation and the domain of non-intentional explanation, because it denies that the relationship is in need of explanation. That relationship, however, is systematic in the way explained. Pluralism is unacceptable, because from the pluralist stance that fact appears as a coincidence. But the systematic correspondence between the two domains cannot possibly be a coincidence—such a coincidence would be truly *miraculous*.[95]

One may object that the relationship between overt action and bodily movements is not as tight as it might seem. Every action of *raising one's hand* may be realised by an event-token of the type *rising of the arm*. But, it seems, many other actions can be realised in various ways. My response to that is twofold. Firstly, it seems clear that many non-basic actions, such as giving a signal or travelling to London, are multiply realisable. However, whether or not *basic*-actions are multiply realisable is not so obvious. No instantiation of raising one's hand, of course, will exactly be like any other instantiation. But that is not to say that raising one's hand is multiply realisable. Rather, that is merely to point out that tokens of that type will, presumably, differ in some respects. Given that, it is not easy to see how basic actions can be multiply realised. But, secondly, nothing of importance depends on that issue. We can grant that all actions—non-basic and basic—can be multiply realised, because certain types of actions will nevertheless be dependent on certain act-neutral types of events. Let me explain.

Consider first the mental antecedents of action. We assume that bodily movements—or act-intrinsic events in general—are caused by neuro-physiological events, and that they can be causally explained by reference to them. That causal dependency supports counterfactuals of the following form.

---

[94] Different metaphors have been used to express that point. MacDonald, for instance, talks about the 'harmony' between mental and physical events—and between the level of mental and physical explanation (1992, p. 231). Cussins and Smith, for instance, say that mental and non-mental theories and entities 'march in step' (Cussins, 1992, and Smith, 1992).

[95] Compare Macdonald, 1992, especially pp. 230-231; Cussins, 1992, pp. 192-200; Smith, 1992, pp. 20-24; Papineau, 1992, pp. 57-58; Stoutland, 1986, p. 46; and Antony, 1991, p. 319.

(CF1) Given relevantly similar circumstances, had the neuro-physiological event *n* of type *N* not occurred, the bodily movement *m* of type *M* would not have occurred.

Further, if an action has been performed for a reason, that particular performance was dependent on the agent's having—or being in—a certain reason-state, and true reason-explanations support counterfactuals of the following from (independently of whether they are construed as causal or non-causal explanations).[96]

(CF2) Given relevantly similar circumstances, had the agent not been in the reason-state *r* of type *R*, she would not have *A*-ed.

If mental states are multiply realisable and, therefore, not type-identical with neuro-physiological states, then it will not be true that the agent's being in *r* depends on the agent's being in *n* (some other neuro-physiological state may realise being in a state of type *R*). But it does not follow that being in a state of type *R* does not depend on being in a neuro-physiological states of a certain type. Presumably, every instantiation of *R* is partly realised by some rather specific state; namely, by a state that partly realises *R*.[97] Let us say that the set $\{N_1, N_2, N_3\ldots\}$ is the set of all realisations of *R*. Given that, counterfactuals of the following form hold.

(CF3) Had none of the n_i's occurred, the agent would not have been in a mental state of type *R*.

Consider now the actions themselves. If actions are multiply realisable, they are, presumably, realised by some rather specific physical event or bodily movement; namely, by a bodily movement that realises the type of action in question. So, let us say that the set of bodily movements $\{M_1, M_2, M_3\ldots\}$ is the set of all the realisations of *A*-ing. Given that, counterfactuals of the following form hold.

(CF4) Had none of the m_i's occurred, the agent would not have *A*-ed.

Actions depend counterfactually on reasons at the level of intentional explanation. At the physical level, bodily movements depend on non-mental antecedents. However, actions depend as well on bodily movements, and the mental antecedents of action—that is, reason-states—depend on their physical realisation. The level of reason-explanation depends, in that sense, on the physical level, even if it is assumed

[96] See, for instance, Schiffer, 1991; Mele, 1992; and Ruben, 2003, who argue that the support of counterfactuals is essential to the explanatory force of reason-explanations.

[97] The neuro-physiological state realises the mental states only partly, if the mental state supervenes partly on the agent's environment—due to environment-dependent content.

that actions and reason-states are multiply realisable. Further, that dependence is systematic in the following sense. Assume that an agent's *A*-ing can be explained by reference to *r*, the agent's being in the reason-state of type *R*. Then, typically, *m*, the physical realisation of that *A*-ing, is dependent on *n*, the physical realisation of *r*, because *m* is caused by *n*.

The fact that the relationship is systematic—the fact that the levels 'march in step' in the sense outlined—cannot be a coincidence, and it is in need of explanation.[98] The causal theory acknowledges that challenge, and different versions of causalism either provide an account of the relationship between reasons for actions and the causes of movements, or they are designed to be in line with some theory—or theories—of the mind that provide the required account.[99] In doing so, the causal theory provides us with an integrated theory of human agency by showing how the 'space of reasons' and the 'space of causes' are correlated.

Non-causalists stress, firstly, that reason-explanations of actions and causal explanations of bodily movements are different in kind, and secondly, that they explain different things. However, both claims are beside the point. Firstly, we saw that causalism is compatible with the second point as it is compatible with the claim that reason-explanations explain actions, whereas neuro-physiological explanations explain bodily movements. Secondly, causalism is compatible with the first claim as it is compatible with the claim that reason-explanations are rationalising explanations that are formulated in intentional terms, whereas causal explanations of bodily movements are not. And thirdly, it became clear that it does not really matter whether actions are identical with, constituted or realised by events. Rather, what matters is that the relationship between them is systematic and not coincidental.

It is only natural, I said, to ask how the efficacy of reason relates to the causal efficacy of the antecedents of bodily movements, given the intimate relationship between actions and bodily movements. I explained then in more detail what that

---

[98] It has been a popular approach to capture that relation of dependence in terms of *supervenience* (see, for instance, Davidson, 1980, essay 11 and 1993 Kim, 1993; Lennon, 1990 and Macdonald, 1992). However, definitions of the relation of supervenience, as it holds between the mental and physical, merely specify the dependence more precisely. That is to say that supervenience is descriptive, rather than explanatory. If supervenience holds, then a specification of that relation tells us in what narrower sense the mental depends on the physical, but it does not explain why it is so dependent. In other words, supervenience is itself in need of explanation—given that it holds. Compare Horgan, 1993, and Kim, 2000, especially pp. 9-15.

[99] For solutions incorporated into a theory of action see, for instance, Davidson, 1980; Goldman, 1970; Lennon, 1990; Mele, 1992; and Pietroski, 2000.

relationship consists in and why it is in need of explanation. To provide an explanation of that relation is not only required for a solution to the mind-body problem and the problem of mental causation, but it is also of central importance to our understanding of rational agency. A comprehensive account of agency must show how the influence of reasons relates to the causes of bodily movements—a comprehensive account must be what I called an *integrated* account. Pluralism does not provide an explanation of that relationship, simply because it denies that it is in need of explanation. But that is why it is unacceptable. Further, insofar as non-causalism is committed to pluralism, non-causalism is unsatisfactory because it fails to provide an integrated account of agency.[100]

Pluralism and Non-Causalism

Pluralism is compatible with both non-causalism and causalism about reason-explanations. Some philosophers think it is undeniable that reason-explanations are causal explanations *and* they reject the project of accounting for the correspondence between the relevant levels of explanation as misguided—they are, in other words, causalists about reason-explanations *and* pluralists.[101]

The kind of causalism that I am defending, however, acknowledges that the relationship between the relevant levels of explanation is in need of explanation. Furthermore, I argued that we should reject non-causalism, because it is committed to pluralism. What remains to be shown is that non-causalism is committed to pluralism.

Reasons, as everyone can agree, influence the actions of rational agents. Non-causalists deny that their influence is causal in kind. Further, they may argue that it is misleading to construe their influence as being directed towards actions—or bodily movements. Rather, reasons influence actions *by* influencing the judgements and choices of rational agents. Their influence is *rational*, and what is influenced is not the action, but the *agent*—the thinking agent or person. To focus on the relation between reasons and action, as non-causalists may insist, obscures that central aspect of rational agency and leads to a misguided quest for a metaphysical connection between reasons and actions.

---

[100] Apparent alternatives are the traditional views of *occasionalism* and Leibinz's theory of *pre-established harmony*. Both views refer to a divine being—God—in order to explain the interactions between the body and the mind. I shall not argue against either view, because I will not engage with the assumption that there is such a being.

[101] See, for instance, Burge, 1993; Baker, 1993; and Putnam, 1999. Compare also Dennett, 1989.

To say that reasons influence actions by rationally influencing their agents sounds very plausible, and it is hard to see why anyone would deny that. However, it is a mistake to think that this observation points, by itself, towards an alternative account of the efficacy of reasons. Everyone can agree with that characterisation of rational agency, simply because it is only a *characterisation* of the agent's ability to choose and act in the light of reasons. More importantly, though, shifting the focus from the efficacy of reasons to the powers of agents merely defers the problem. Given that an agent's bodily movements are caused by non-mental states and events, and given the intimate relationship between actions and movements, the agential power to act in the light of reasons is itself in need of explanation. In the first chapter we distinguished between four different positions in the metaphysics of agency, which provide different accounts of an agent's power to act for reasons: the reductive standard-causal model, the non-reductive agent-causal model, volitionism and pluralism.

I rejected the non-reductive approach and volitionism in the first chapter. The remaining options are, then, the reductive model and pluralism. At the beginning of this chapter I pointed out that causalism about reason-explanations goes hand in hand with the reductive standard-causal approach. Given that, the remaining option for non-causalism is pluralism. And given, furthermore, that the four named positions exhaust the viable possibilities, non-causalism is committed to pluralism.

One may object that it has not been shown that non-causalism about reason-explanation is incompatible with the standard-causal model of agency. One can envision a view that says that agent-involving mental states and events cause and causally explain actions, and denies that reason-explanations are causal explanations. Stoutland's later position, discussed above, is of that kind.[102] I argued, though, that intentionalists—proponents of that view—in fact have good reason to endorse causalism about reasons and reason-explanations. Another version of that view can be envisaged in combination with externalism about reasons (the view that reasons are facts, rather than mental attitudes or their contents).[103] One may hold that agent-involving mental states and events cause and causally explain actions and deny that reason-explanations are causal explanations, because mental states and events are not

[102] See *Agreement between Causalists and Non-Causalist?*, pp. 89.
[103] See *Externalism About Reasons*, pp. 67.

reasons for action at all. I argued, though, that externalism about reasons is compatible with causalism and that there is no reason to deny that explanation in terms of mental states and events are reason-explanations, given that they stand in some appropriate relation with the relevant reason-facts. Further, I showed that explanations in terms of reason-facts are not distinct from explanations in terms of mental states and events, since the former are merely condensed versions of the latter.

Those considerations do not show conclusively that non-causalism about reason-explanation and the standard-causal model of agency are incompatible, but they strongly suggest so. Most non-causalists, however, would not want to combine their view with the standard-causal model anyway. They commit themselves to the non-reductive model, to volitionism, or to pluralism. Non-causalism is committed to pluralism insofar, and only insofar, as reductionism, non-reductionism and volitionism are dismissed as options for non-causalism, leaving pluralism as the only option available. I argued, though, that not even pluralism is a *viable* option as it fails to provide an integrated account of agency.

Remaining Issues and Problems

Note, first of all, that the case for causalism is not yet complete. Causalism, as I understand it, acknowledges that the relationship between the level of reason-explanation and the level of non-intentional explanation is in need of explanation. What remains to be shown, of course, is that the causal theory can explain the relationship. In the following chapter, I will suggest an account of that relationship, which is compatible with non-reductive physicalism and which presupposes causalism about reason-explanations.

Secondly, there are still important questions to be answered with respect to the offered causal model of acting for reasons. I pointed out, already in the first chapter, that reason-states must cause actions *in the right way*. In particular, it must be excluded that they cause them by way of so-called deviant or wayward causal chains. I will return to that issue in chapter 4.[104]

Thirdly, according to causalism an action can be explained in terms of reason-states, even if it was not performed for *good* or *normative* reasons. The question is,

[104] Chapter 4, pp. 175.

though, whether the agent must, somehow, *treat* or *take* some consideration *as* a reason in order to act *for* that reason, or whether it is sufficient that the performance of the action can be rationalised in the light of some of the agent's attitudes.

Finally, there is one objection to the causal approach closely related to the previous point, which deserves to be taken seriously. That objection goes, roughly, as follows. The possibility of deviant causal chains highlights a fundamental problem for the causal theory. The theory says that reason-states both rationalise *and* cause actions, and it requires that those two aspects go hand in hand—it requires that reason-states cause an action if and only if they rationalise it. But that misses a very important feature of acting for reason. Namely, that an agent who acts for a reasons is motivated to perform the action *in virtue of* having that reason—in virtue of recognising that there is reason to do it. That means, in terms of the causal theory, that the reason-states should not *merely* cause *and* rationalise the action, but should cause the action *in virtue of* rationalising it. Reason-states would need to be *rational causes*, rather than states that cause and rationalise.[105] I will address that objection together with the previous question in chapter 4.[106]

[105] See, for instance, Antony, 1989, Brewer, 1995 and Wedgwood, *forthcoming*. Antony argues against Davidson's model that it does not explain '*how* reasons cause actions so as to rationalize them simultaneously. The explanatory force of rationalizations is partly explained by the fact that reasons are things that have causal efficacy, but we also need to know how it is that reasons can have efficacy *in virtue of their reasonableness*' (p. 168).

[106] Concerning causation in virtue of rationalising see chapter 4, pp. 183. Concerning the issue of treating as a reason see chapter 4, pp. 193 and compare note 59 in this chapter, p. 89,

Chapter Three: Causalism as Non-reductive Physicalism

In the previous chapter I argued that a satisfying account of agency must be an integrated account: it must provide an account of the relation between reasons for action and the causes of movements. The systematic relationship between actions and bodily movements—the fact that they 'march in step'—cannot be coincidental; it *is* in need of explanation. Let us call the problem to provide that explanation the *coincidence problem*. Pluralism does not solve the coincidence problem, because it does not acknowledge it as a problem. Rather, it dismisses as misguided the project to relate the causes of movements with reasons for actions. I argued that we should reject pluralist positions—hence, non-causalism—precisely because of that. However, the mere fact that the causalist position does acknowledge the problem, does not give us reason to prefer it to non-causalism. What is needed is reason to believe that the coincidence problem can be solved within a causalist framework.

Non-causalists, however, may argue that the project to solve the coincidence problem is bound to failure. A convincing argument to that conclusion would show that the project to relate the causes of movements with reasons for actions is indeed misguided, and it would show that the present issue does not give us any reason to prefer causalism. A first argument of that kind goes along the following lines.

(1) The relationship between the causes of movement and reasons for action is in need of explanation.
(2) An explanation of that relationship requires either type-identities between mental and physical types or a reduction of psychology.
(3) Type-identity theories are untenable.
(4) Reason-explanations are irreducibly formulated in intentional terms. That is, psychology is not reducible.
(5) Hence, causalism either fails to account for the relationship, or it is committed to either one of two unfavourable positions; namely, the type-identity theory or some form of reductionism.

Another argument goes as follows.

(1') Every version of causalism faces the problem of causal exclusion.
(2') Only type-identity and reductive theories can solve that problem satisfyingly.

(3') Type-identity theories are untenable.

(4') Reason-explanations are irreducible formulated in intentional terms. That is, psychology is not reducible.

(5') Hence, causalism either does not solve the problem of causal exclusion or it is committed to either one of two unfavourable positions; namely, the type-identity theory or some form of reductionism.

In response to the first argument I will argue that an explanation of the relationship neither presupposes type-identities nor reduction; in other words, I will reject assumption (2). My response to the second argument will question assumptions (1') and (2'). Before that I will say more about the layered model of the world, which is usually presupposed as a metaphysical framework in debates concerning the present issues, and I will say more about the third and fourth assumption of the two arguments.

Causalism comes in different versions. In the following, I will consider the identity theory, functionalism, and non-reductive physicalism.[1] My main aim is to argue *for* causalism *in general*. What needs to be shown in order to complete the argument of the previous chapter is that causalism can solve the coincidence problem. The most general distinction with respect to causalism is between reductive and non-reductive versions—that is, between versions of causalism according to which intentional psychology is reducible and ones that deny that. If psychology is reducible, the coincidence problem, as I will show, does not arise. However, if psychology is not reducible, the problem does arise and it must be shown how non-reductive theories can solve it. Some causalists argue, or insist, that psychology is reducible. But that is not the strategy that I shall pursue. Rather, I will argue that, as far as the coincidence problem is concerned, the tenability of causalism does not depend on the reducibility of psychology.

My second aim is to contribute to a defence of non-reductive physicalism, which is my preferred version of causalism. To show that non-reductive theories can solve the coincidence is the first part of that. In a second part, I argue that the so-called causal exclusion argument is not decisive against non-reductive physicalism.

[1] Those theories are, of course, standard theories of the mind; in particular, they are standard accounts of the relationship between the body and the mind. It should be clear—at least, it should in the following become clear—why those theories are the relevant versions of causalism.

Levels of Explanation

We are interested in the relation between the causes of bodily movements and reasons for actions. In recent philosophy of mind, issues of that kind have been approached by focusing on the relation between different *levels of explanation*.[2] According to that 'layered model', as Kim calls it, the world is 'stratified' into different orders, whereby each layer or level is associated with a certain theory (or the conjunction of theories of a certain kind), which ranges over certain entities and properties. Each level is a higher- or a lower-level in relation to other levels.[3] The theories, entities and properties, which are characteristic of a given level, can accordingly be called higher- or lower-level theories, entities and properties. Further, we can talk about higher- or lower-level predicates, regularities and explanations, which are characteristic of the higher- and lower-level theories in question.

In virtue of what is one level higher or lower in relation to another? The criterion offered by Kim is *composition*. If the entities described at the level L are entirely composed by entities described at some other level L', then L is a higher-level of explanation in relation to L', and L' is a lower-level in relation to L. The entities of higher-levels are, as Kim says, 'exhaustively decomposable' into lower-level entities.[4]

Sometimes it is pointed out that levels of explanation differ with respect to the degree or extent of *abstraction*. Describing an entity and explaining its behaviour at a higher-level, we abstract from its composition: we abstract from how it is composed and what it is composed of—in other words, we abstract from what parts it is composed of and how those parts are related to each other. In relation to the lower-level theory, which describes the *parts* and the interactions between them, the higher-level theory describes and explains the behaviour of the entity as a *whole*.

Given that model of levels of explanation, the relation between reasons for actions and the causes of movements can be construed as follows. A reason-explanation of an action, we assume, cites those mental states of the agent that rationalise the performance of that action. Mental states are characterised and referred to by a theory that employs intentional vocabulary, and they are attributed to agents.

---

[2] See, for instance, Owens, 1989, and Kim, 2000, pp. 15-19.
[3] Here we can ignore the question whether there is, or has to be, a *lowest*—or *highest*—level of explanation. To distinguish levels relationally is sufficient for present purposes.
[4] Kim, 2000, p. 15.

That theory—call it psychology—is about the behaviour of the agent as a whole. It abstracts from facts about the constitution or composition of the agent, and from facts about the physical—or neuro-physiological—causes of the agent's movements. In relation to a theory that is about the causes of the agent's bodily movements, psychology is a higher-level theory that describes and explains higher-level patterns of behaviour.

Appeal to composition is used to classify different levels of explanations as higher- and lower-levels. The claim that higher-level entities are entirely decomposable into lower-level entities does, obviously, tell us something about the relation between higher- and lower-level *entities*. But what is more interesting, and more controversial, is how the relationship between higher- and lower-level *properties* and *theories* should be construed.

On a standard model of scientific explanation, the best account of those relations is provided by means of scientific reduction.[5] On that model, a higher-level theory is reducible to a lower-level theory only if there are so-called bridge laws that relate the higher-level properties with lower-level properties such that the higher-level regularities are deducible from the lower-level regularities (in conjunction with the bridge laws).[6] Some argue that such bridge laws cannot have the form of biconditionals. The reason is, simply, that a biconditional does not provide an *account* of the relation between the levels. A biconditional between mental and physical properties, for instance, does not explain why a given mental property is instantiated whenever a certain physical property is instantiated.[7] In other words, the biconditional merely reaffirms what we were assuming all along; namely, that there is some systematic relationship between the two domains. On the basis of that, one may argue that bridge laws must have the form of identity statements. Identity between the higher- and lower-level kinds would not merely claim a systematic relationship between the two levels, but it would *explain* why that relation holds. If the higher- and lower-level kinds are identical, the question why a given higher-level property is

[5] A standard reference is Ernest Nagel's *The Structure of Science*, 1961.
[6] Note that the relation between higher- and lower-levels is asymmetric in the following three ways: the entities of the higher-level are composed by lower-level entities, the higher-level is an abstraction from lower-levels, and the higher-level regularities can be deduced from the lower-level laws (in conjunction with the bridge laws). A further asymmetry is that the lower-level theory is explanatory of the higher-level theory, which is why the reduction of a higher-level theory can be regarded as a *vindication* of that theory.
[7] Compare, for instance, Kim, 2000, chapter 4.

instantiated whenever a certain lower-level property is instantiated has a simple answer: *because* instantiations of the higher-level property *are* instantiations of the lower-level property. It seems, then, that the two most important questions are whether mental kinds are identical with lower-level kinds, and whether intentional psychology can be reduced to some non-intentional lower-level theory.

However, I shall assume, with the opponent, that mental kinds are not identical with lower-level kinds, and that intentional psychology—and hence, reason-explanation—is irreducible. Given that, the question we have to ask is whether there are alternative accounts of the relation between the levels of explanation in question. To begin with, let us consider what a reduction of psychology *would* give us.

Firstly, reduction solves the coincidence problem; it would provide an account of the systematic relation between the two domains and it would, thereby, explain why they march in step. Secondly, reduction avoids the problem of causal exclusion. If the higher-level kinds can be reduced to, and if the higher-level laws can be deduced from the lower-level laws, then there is no reason to think that the higher-level entails or presupposes irreducible causal powers that compete with the causal powers described by the lower-level. And thirdly, the reduction of the higher-level theory is explanatory—it explains why the claims made and entailed by the higher-level theory hold. To learn that water *is* H_2O and that regularities that describe the properties of water can be reduced to chemistry is not just to gain another way of talking about water. Far from it, by identifying water with H_2O we gain understanding of why, for instance, water interacts with other substances in the way it does, or why it shares some properties with some other substances, but not with others, and so forth.

A reduction of a higher-level theory is worth wanting because it solves—or, rather, avoids—the coincidence problem and the problem of causal exclusion, and because it explains and vindicates the higher-level theory. What is needed in defence of non-reductive physicalism, however, is only the former. The important point here is that we can distinguish between an explanation of the fact that the levels march in step and an explanation of the higher-order theory in lower-level terms. Given that, we can separate the question of whether the coincidence problem can be solved from the question whether the higher-level theory in question can be reduced. A solution to the coincidence problem explains why the levels march in step, whereas a successful

reduction explains—in addition—the higher-level theory in terms of a lower-level theory.

Now I will briefly turn to reductive theories, their shortcomings, and I will present the motivations behind non-reductive physicalism. Then I will propose an alternative solution to the coincidence problem that is committed to neither type-identities nor reductionism. And in the last section of this chapter I will offer a response to the causal exclusion argument.

The Type-Identity Theory

According to the type-identity theory, mental kinds (types or properties) are identical with lower-level kinds (types or properties).[8] If identity holds between types, there *will be* bridge laws that identify higher-level kinds with lower-level kinds, and the higher-level mental theory *will be* reducible to some lower-level theory.[9] This would give us straightforward solutions to both the coincidence problem and to the problem of mental causation. We can even say that these problems do not arise for the identity theory in the first place, simply because it says that mental kinds *are* lower-level kinds.

According to mainstream opinion, however, the identity theory is untenable. Arguments from multiple realisation and from considerations concerning the modality of identity are taken to be decisive against it.[10] Only recently there have been attempts

[8] It is common to assume that it does not make a significant difference whether the claimed identity holds between states, properties or kinds. Note further that it is not obvious what the appropriate lower-level theory is supposed to be. Is it physics, chemistry, neuro-physiology? In a recent defence of the identity theory, Thomas Polger talks at one point about identity between mental and *biological kinds* (see Polger, 2004, p. 136). In another passage he talks about identity between mental and *physical states* (p. 35). Polger assumes, apparently, type-identity all the way down: from mental to physical properties *via* neuro-biological (and other lower-level) properties.

[9] Some philosophers think that reducibility stands and falls with type-identity. Others, however, would reject the claim that mental types are identical with physical types if and only if psychology is reducible to physics (presumably by way of reducing psychology to neuroscience, and reducing neuroscience by some further steps to physics) as too strong. They argue that type-identity is not necessary for reduction, since bridge laws featuring biconditionals, rather than identities, are sufficient. It seems safe to say, however, that type-identities are a sufficient condition for reduction.

[10] The argument from multiple realisation is due to Hilary Putnam, 1967. It says, very roughly, that mental types cannot be identical with lower-level types, since it is possible that they are realised in different ways: different individuals (of different species or kinds) may be in a mental state of the same kind without being in a physical state of the same kind (or without being in any physical state at all). The argument from the modality of identity is due to Saul Kripke, 1972. It says, very roughly, that the apparent contingency of the relationship between mental states and brain states is incompatible with the necessity of identity. For a recent discussion of both arguments see Polger, 2004.

to revive the identity theory.[11] What counts in favour of the theory is that it provides a very neat solution to the mind-body problem in general, and to the problem of mental causation in particular. And if one thinks that the type-identity solution is the only satisfying solution to the problem of mental causation, one has good reason to seek responses to the arguments from multiple realisation and modality.

However, the identity theory faces another objection. Namely, that it cannot do justice to the nature of intentional reason-explanations, since it entails the reducibility of psychological explanations to non-psychological ones. An essential feature of reason-explanations of action is that the performance of the action is rationalised and justified by referring to the agent's reasons under their intentional description. That is to say that rationalisation and justification of action is essentially provided in intentional terms; mental states rationalise or justify actions only under their intentional descriptions. The identity theory does—and cannot—do justice to this fact, so the objection goes, since it entails that intentional psychology is reducible to non-intentional theories.

According to the outlined standard account of scientific reduction, a reduction of psychology to some non-intentional theory entails that the principles and regularities that support intentional explanations and predictions can be deduced from non-intentional lower-level regularities (in conjunction with bridge laws). That would mean that everything that can be explained by psychology could, in principle, be explained by the lower-level theory. And since the lower-level theory does not employ intentional vocabulary, it would mean that everything that can be explained in intentional terms could, in principle, be explained in non-intentional terms, which is incompatible with the fact that reason-explanations are essentially formulated in intentional terms.

Further, it would mean that the regularities that ground reason-explanations can be deduced from non-intentional laws, and that the latter are explanatory of the former. But reason-explanations *rationalise* actions, and they are, presumably, based on normative principles. It is simply very difficult to see how normative principles of

[11] See for instance Polger, 2004.

rationality could be deduced from and *explained* by non-normative lower-level laws formulated in non-intentional vocabulary.[12]

The two main points of the objection are the following. Reduction is implausible firstly because reason-explanations are essentially formulated in intentional terms, and, secondly, because the norms that ground those explanations are normative principles that can neither be deduced from nor explained by non-intentional lower-level theories—the norms of rationality have, as Davidson has put it, *no echo* in the physical domain.[13]

Closely connected to those two points is the following. The reduction of psychology means not only that psychological regularities can be deduced from and explained by a non-intentional theory. But it means also that psychological regularities can, at least in principle, be *discovered* without doing psychology. And there is no reason to think that we could discover psychological laws and the norms of rationality by doing, say, neuro-physiology. The three basic components of reduction are bridge laws of some sort, deducibility of the higher-level laws, and the explanatory constraint—the constraint that the lower-level theory must be explanatory of the reduced higher-level theory. Ned Block, for instance, has pointed out that it is important to distinguish the deducibility from the explanatory constraint, and that the former does not entail the latter. Block says that

> if one has to do *psychology* to discover basic laws of *physics*, the deduction of [the] laws of psychology from their images in physics won't be as explanatory as one might wish.[14]

What Block is suggesting, it seems, is that the explanatory constraint should accommodate epistemological issues concerning the discovery of certain regularities. What is pointed out, in other words, is that reduction concerns not only metaphysical relations between kinds or properties and logical relations of deducibility. A reduction

---

[12] One may appeal to *a priori* considerations of that sort in order to argue for the stronger claim that a reduction of psychology is *impossible*. For the present purposes, however, the weaker claim that reduction is *implausible* is strong enough to support the dialectic of my considerations.

[13] Davidson, 1980, p. 231.

[14] Block, 1997, p. 111. Note that the anti-reductionist point that Block makes is stronger than the one that I have promoted. Block's suggestion is not simply that we have to do psychology in order to discover psychological regularities, but that we have to do psychology in order to discover lower-level laws. In order to support that claim, one might argue that in order to discover the multiple realisations of mental properties we have to consult psychological investigation. Or one might argue that scientists can discover the mechanisms that implement certain mental processes only by investigating the brains of *conscious* beings who are able to follow instructions and to interact with their environment *intentionally* and *rationally*.

is supposed to be explanatory of the higher-level theory, and, according to Block, that cannot be equated with the deducibility of the higher-order laws. That means that even if psychological regularities—including the norms of rationality—were deducible, there would still be further issues concerning the discovery of higher- and lower-level laws that would give us reason to question the reducibility of psychology.

Functionalism

According to Jackson and Pettit, functionalism is a view about the truth conditions for the attribution of mental attitudes. It says, roughly, that an agent *S* is in a mental state of type *M* if and only if *S* is in some lower-level state that plays the causal role characteristic of *M*.[15] Functionalism, one might add, is about how mental states are to be individuated; namely, solely in terms of their functional—that is, causal—roles.[16]

For further characterisation I will use the common distinction between *role* functionalism and *realiser* functionalism.[17] According to role functionalism, mental kinds are identical with functional—causal role—kinds. On that view, mental properties are second-order properties, which are multiply realisable by all first-order properties that satisfy the functional role specified by the second-order property.[18] Realizer functionalism is about mental concepts (or predicates), rather than properties (types or kinds). It says that mental concepts can be analysed or characterised by specifying their functional roles. On that view, there are no general mental kinds, but only species- or system-relative mental kinds. What unifies the diverse first-order realizations is not that they instantiate the same second-order property, but that they fall under the same functional concept. So, for a given species or type or system, mental states can be identified with the first-order states that fall under the mental concept (in individuals of that species or type of system).

---

[15] Compare, for instance Jackson and Pettit, 1988, p. 384. Two classic sources are Putnam, 1967, and Lewis, 1980. For further discussion see, for instance, Block, 1978; David, 1997; Polger, 2004; and Shoemaker, 1981.

[16] To specify the functional or causal role of a mental state is to specify its typical causes and effects. A typical cause of pain, to use a common example, is tissue damage and a typical effect is groaning or wincing. Being in pain is then to be in whatever state that satisfies the functional role of pain—in whatever state that has the causes and effects characteristic of being in pain. In the case of human beings that state is, presumably, some neurological state (or brain state).

[17] Compare Jackson and Pettit, 1988, p. 385.

[18] Generally, a second-order property is the property of having a first-order property that satisfies the functional role that is characteristic of the second-order property. For instance, for human beings to be in pain is to be in the first-order neurological state that satisfies the functional role that is characteristic of the second-order property of being in pain. Compare, for instance, Kim, 2000, p. 19.

Is functionalism a reductive theory? Beginning with role functionalism, we have to distinguish two claims. Role functionalism is a form of reductionism in the sense that it identifies mental kinds with functional kinds, and in the sense that it assumes that psychology is reducible to a functional theory. That is, role functionalism is committed to the claim that psychological explanations can be reduced to functional explanations. There is, however, the further question of whether mental—second-order—properties can be identified with or reduced to the disjunctions of first-order properties that realise them. Role functionalism is not committed to reductionism in that sense.

Realizer functionalism is reductive in the sense that it reduces mental concepts and explanations to functional concepts and explanations, and in the sense that it identifies mental states with first-order realizations relative to a given species or type of system.[19]

According to both kinds of functionalism, mental concepts can be reduced to functional concepts. Role functionalism goes beyond that by making claims about mental kinds, rather than just concepts. According to functionalism, then, for each intentional description of a mental state there is an equivalent functional description of that state, and intentional explanations can be reduced to explanations in terms of functional states and roles. Given that, it is clear that functionalism is subject to the anti-reductionist considerations presented in the previous section. A functionalist reduction of psychology is implausible, because it is incompatible with the fact that reason-explanations are essentially intentional explanations, and because it entails that the normative principles of rationality that ground reason-explanations can be deduced from and explained by non-intentional—in the present case, functional—theories (in conjunction with bridge laws).[20]

In the next section I will turn to non-reductive physicalism. Before that, however, it will be instructive to see how functionalism solves the coincidence problem. Role

[19] Note that there are two different notions of reduction employed. The former can be called ontological, the latter semantic reduction. Role functionalism assumes that there is not only talk about the mental, but that there are mental kinds. A reductive version of the theory must employ an ontological conception of reduction that reduces mental to non-mental kinds or properties. Realizer functionalism, however, restricts itself to mental concepts and does not stipulate the existence of kinds or properties. For such a theory, reduction is merely a semantic matter. To reduce the mental is to establish equivalence relations between mental and non-mental concepts, descriptions and sentences. On the two kinds of reduction see for instance Schiffer, 1987, p. 142.

[20] Compare, for instance, Kathleen Lennon, 1990, chapter 6. Lennon argues that a functionalist reduction is implausible because it entails non-intentional descriptions of essentially intentional facts.

functionalism identifies mental properties with functional second-order properties (that is, relational properties concerning the causal relations between first-order properties). Any system or agent whose first-order properties satisfy—or occupy—the causal roles specified by the second-order property realises —and thereby *has*—the mental property in question. The solution to the coincidence problem is straightforward. What explains that the two levels of explanation march in step is the fact that the mental second-order property is realised by the first-order property—in the outlined functionalist sense of realisation. Since having the right first-order property *is* having the mental second-order property, it is no coincidence that reason-explanations, which are in terms of second-order properties, and causal explanations of movements, which are in terms of first-order properties, march in step. Similar considerations apply to realizer functionalism.[21]

Non-reductive Physicalism

Non-reductive physicalism is the conjunction of the claim that psychology is not reducible, and physicalism. Physicalism, one may think, says that everything is physical. It is common to distinguish physicalism from claims such as that there are only material objects or that everything is composed of matter. Physicalism refers to physical things, rather than material ones, and it leaves it up to the physical sciences to specify what counts as a physical entity. Further, the view concerns not only objects and what they are composed of, but also events and properties. Now, if non-reductive physicalism is to be a consistent position, then physicalism cannot be construed as the view that all objects, events and properties are physical. Recall that if psychology is not reducible, then mental properties cannot be identified with non-mental properties. Non-reductive theories of the mind say that there are mental properties that are not identical with physical properties. Hence, they are obviously

[21] As with respect to the problem of mental causation, it is commonly accepted that realiser functionalism is not subject to the causal exclusion problem. Since mental states are identified with the first-order realizations, there is no problem of causal overdetermination. In fact, that is thought to count in favour of realiser functionalism (as opposed to role functionalism). Role functionalism identifies mental properties with second-order—that is, higher-level—properties. These properties are thought to figure in causal explanations, which presupposes that we think of them as being causally relevant. But each token of a mental property is realized by some first-order property, and it seems plausible to say that it is the realization that is doing the causal work. One may think, therefore, that role functionalism has a problem with mental causation: either mental properties are merely explanatory (that is, not efficacious), or, if we assume that they are causally efficacious, they merely overdetermine effects that already have a sufficient cause (namely the first-order realizations of the mental states in question).

incompatible with a view that says that all properties are physical. If it is to be compatible with non-reductionism, physicalism must be specified accordingly. A common suggestion goes along the following lines. Physicalism says, firstly, that all concrete objects are composed of, or constituted by, physical entities, and, secondly, that all properties and events are *dependent* on physical properties and physical events.[22] Physicalism says, in other words, that physical particulars and properties are the *basic* or *fundamental* particulars and properties.[23]

We can, then, characterise non-reductive physicalism as the conjunction of physicalism and the claim that psychology is not reducible. In particular, non-reductive physicalism rejects the claim that psychology can be reduced to a functional theory.[24] Functionalism assumes three different levels of explanation: the higher-level of psychological explanation, the lower-level of physical explanation, and an intermediate level of functional explanation.[25] Functionalism claims that the higher-level can be reduced to the intermediate functional level.[26] Non-reductive physicalism denies that. It says that psychological explanations are essentially and irreducibly intentional explanations.

The claim, however, that intentional psychology cannot be reduced to an intermediate functional theory does *not* entail there is no non-redundant intermediate functional theory. In the following I will suggest a solution to the coincidence

---

[22] Compare, for instance, Beckermann, 1992, pp. 1-2, and Crane, 1995, pp. 211-212. Compare also chapter 2, p. 106, note 98.

[23] That is the standard account of non-reductive physicalism. An alternative account goes as follows. Arguably, we can distinguish between two different conceptions of reduction. On the first, reduction is an ontological issue insofar as it concerns the relation between mental and physical *properties*. On the second conception, reduction concerns the relation between *theories* (their vocabularies and explanatory powers). Non-reductive physicalism is clearly committed to non-reductionism in the second sense. It is not clear, though, whether it is also committed to ontological reductionism. If it is not, then it is not committed to the claim that there are irreducible mental properties. In that case, the view says only that the psychology is not reducible (that psychological concepts and regularities are not reducible and that psychological explanations are *genuine*). I can agree that reduction is, first and foremost, a relation between theories. The question, however, whether the reduction of a theory presupposes ontological reduction is beyond the scope of this work. I will assume, throughout, that the standard construal of non-reductive physicalism is correct.

[24] Depending on what conception of reduction is employed, that claim entails either that mental *properties* are not reducible to functional properties or that statements involving mental *concepts* are not equivalent to (and not reducible to) functional concepts.

[25] We can ignore the fact that the lower-level consists itself of different levels.

[26] Compare, for instance, Fodor, 1991. Fodor says that according to functionalism, reason-explanations '*reduce* to explanations articulated in terms of functional states [because] beliefs and desires *are* functional states. And, for each (true) psychological explanation, there will be a corresponding story, to be told in hard-core science terms, about how the functional states that it postulates are 'realized' in the system under study' (p. 29).

problem for non-reductive physicalism that assumes that there is such a theory, and that this theory can explain why the higher-level of intentional psychology and the non-intentional lower-level theory march in step.

The Coincidence Problem

In order to solve the coincidence problem one must provide an account of the systematic relation between the two levels of explanation (namely, the level of intentional reason-explanation, and the level that provides explanations of bodily movements in non-intentional terms)—one must explain why it is no coincidence that they march in step. Bridge laws that relate mental with physical types or a functional reduction of mental properties (or concepts) would provide that explanation. But non-reductive physicalism is incompatible with both type-identities and functionalist reduction. Is there an alternative way to solve the coincidence problem?

Antony and Levine, for instance, argue that a defence of non-reductive physicalism must provide a 'realisation theory' that shows why a system—or agent—endowed with the mechanisms described at the lower-level meets the constraints described at the intentional level, and how the mechanisms mediate the higher-level causal regularities.[27] But how can that be achieved? What is a 'realisation theory'?

Adrian Cussins says that in some cases of physical realization we can *see*—grasp or perceive—a relation of *coherence* between what is described at different levels, if we can *see* how the structures and mechanisms described at the lower-level realize or implement the structures described at the higher-level. For illustration Cussins considers how we can see that the mechanism of the heart realises the function of circulating blood. In cases like that, one does not need to know any bridge laws in order to see why it is that an organ with that organisation—why a mechanism of that sort—realises or implements the function of circulating blood. And one does not need to know bridge laws in order to see the coherence between what is described at the functional level and the level of realisation. In such cases, the relation between the

[27] Compare Antony & Levine, 1997. In a similar vein Botterill & Carruthers say that it must be rendered 'unmysterious that a given special-science law obtains, given the ways in which the properties involved in that law can be realised in physical mechanisms, and given the lower-level laws which govern that lower level.' (p. 187) See also Fodor, 1989; Segal & Sober, 1990; and Antony (1995), who all claim that higher-level laws must be 'mediated by' underlying mechanisms. On the related point that physical realisation can *explain* supervenience, compare Kim, 2000, especially pp. 19-24.

levels of explanation is, as Cussins says, *intelligible*. We can comprehend why the underlying mechanism realises what is described at the higher level. He says that

> [where] there is an intelligible connection between two levels of discourse there need be no laws governing the relation between the levels, but only this constraint: that a person who understands both levels of discourse understands why it is that having the structure given in the lower level is a way of having the function given in the upper level.[28]

Cussins, of course, does not think that in the case of the mental we will simply be able to *perceive* such an intelligible connection between the levels. The situation is, obviously, far more complex than it is in the case of an organ's function and its anatomical organisation. What Cussins envisages for the case of the mental is a 'scientific theory' that mediates between the two levels—a theory, which is about the relation between the two levels.

In a similar vein, Horgan and Woodward argue that neither bridge laws nor reduction is necessary in order to vindicate common sense psychology. What needs to be shown, rather, is that the lower-level mechanisms, which realize mental states and regularities, 'preserve the causal architecture' of intentional psychology.[29]

Those suggestions are in line with each other. In order to solve the coincidence problem, we need to understand why and how the lower-level realises the higher-level. And a theory that shows that the mechanisms described at the lower-level preserve the causal architecture of the higher-level seems to do that, because the fact that the lower-level mechanisms preserve the causal architecture of psychology would explain why they march in step.

It is important to note that those suggestions have some significant features in common with functionalism. For what functionalism does, amongst other things, is to show how the lower-level preserves the causal architecture of the higher-level by showing that the lower-level mechanisms realise the causal structures described by psychology.[30] Further, to abstract from the nature of the entities that actually stand in the specified causal relations, and to focus on the causal architecture of a system is precisely the functionalist strategy. It seems that a theory that shows how the lower-

[28] Cussins, 1992, p. 204.
[29] Horgan & Woodward, 1985, especially pp. 219-224.
[30] When discussing the example of the heart, Cussins talks about the organs realizing the *function* of pumping blood, and he seems to suggest that a theory that explains the relation between different levels must show 'why it is that having the structure given in the lower level is a way of having the *function* given in the upper level' (p. 204, my emphasis).

level mechanisms preserve the causal architecture of the higher-level would be a functional theory.

But functionalism is a reductive theory—how can it possibly be compatible with non-reductive physicalism? It is one thing to claim that psychology has a causal structure or architecture that can be *described* by a functional theory, which abstracts from the nature of the states and events that actually stand in the described causal relations. It is quite another thing, though, to claim that psychology can be *reduced* to that theory.

Recall that a successful reduction must be explanatory. To reduce some higher-order theory is to explain why the higher-order laws hold and how they relate to the lower-level laws. Non-reductive physicalism denies that psychology is reducible, mainly because it denies that any non-intentional theory can be explanatory of the intentional regularities described by psychology. Non-reductive physicalism is, therefore, incompatible with functionalism. But it does not follow that it is incompatible with a theory, which describes the causal structure or architecture of intentional systems in functional terms.[31] In fact, there is no obvious reason why non-reductive physicalists should deny that such a functional theory is possible.

Given all that, we can see that non-reductive physicalism can solve the coincidence in the same way as functionalism. What is crucial to the functionalist solution is not that functional properties (or concepts) are explanatory of mental ones, but that the lower-level mechanisms are shown to be a realisation of the causal structure described by psychology. And that is all that is needed in order to explain why the two levels march in step—without presupposing reduction, type-identities or bridge laws.

The challenge that non-reductive physicalism cannot solve the coincidence problem stems from the assumption that a solution of that problem requires either type-identity or reduction. Now we can see why that is mistaken. Functionalism is a reductive theory. But the solution to the coincidence problem provided by functionalism does not depend on that. For that reason the functionalist solution is available to non-reductive theories as well. The mistake is to think that arguments

[31] Horgan and Woodward talk about the causal architecture of the higher-level *theory*. Strictly speaking, it is of course only what is described by that theory that *has* a causal structure. But we can talk of the causal structure or architecture of a theory in the sense that the causal relations described by the theory and the causal explanation provided by it exhibit a certain structure or architecture.

against functionalist reduction are also arguments against a functional characterisation of the causal architecture of psychology.

The Causal Exclusion Argument

I will now turn to the so-called causal exclusion argument against non-reductive physicalism. Many philosophers think that this argument is a serious problem for non-reductive theories of the mind—some think that it is decisive against them. Firstly, I will outline the exclusion argument. Then I will distinguish between three versions of the argument that address three different versions of non-reductive physicalism, and I will argue that the causal exclusion argument is not decisive against any of the three versions. According to non-reductive physicalism, mental events are dependent on physical events. Causal exclusion and overdetermination, however, require distinct and independent causes. I will argue that the burden of proof lies with the opponents of non-reductive physicalism, who have to explain how metaphysically *dependent* events can possibly overdetermine an effect or exclude each other from being causally efficacious.

The causal exclusion argument has been formulated in different ways. A first difference concerns the kinds of entities that are said to exclude each other. On some formulations the exclusion concerns mental and physical *properties*, others talk about mental and physical *events*, and some formulate the argument simply in terms of mental and physical *causes* and *effects*.[32] A second relevant difference concerns the mode of causal exclusion. Some philosophers insist that what is at stake is the causal *efficacy* of mental events or properties. Others, however, say that the efficacy of physical events excludes the causal *relevance* of mental events or properties.

What is common and central to all formulations of the argument, though, is the following intuition concerning causal exclusion and causal overdetermination. Suppose that c is sufficient to cause the occurrence of e, and that e has another cause c^*, which is distinct from and not part of c. Let us say that an event e_1 is a *sufficient* cause of the event e_2, if e_1's occurrence is, in the circumstances, sufficient for the occurrence of e_2. And e_1 is a *partial* cause of e_2, if e_1 is *a* cause of e_2 in the sense that

[32] Formulations in terms of causes and effects can be found, for instance, in Lowe, 2003a, and Merricks, 2001. For formulations in terms of instantiations of properties compare Kim, 1993 and 2000, Crane, 1995 and Menzies, 2001. Kim and Yablo, 1992, think that it is of no significance whether the argument is put in terms of causes, events or property instantiations.

e_1 is itself not sufficient for the occurrence of e_2, but it is part of a complex event, which is sufficient for the occurrence of e_2. Now, c^*, the additional cause of e, is either a sufficient or a partial cause of e. If c^* is sufficient, then c and c^* exclude each other from being *the* cause of e, because they overdetermine its occurrence. If c^* is a partial cause of e, then c excludes c^* from being causally efficacious, since c is already causally sufficient for the occurrence of e. (I will call causes that exclude each other or overdetermine their effects in that sense, *rival* causes).

These intuitions concerning exclusion and overdetermination are considerably strong and straightforward *only insofar* as they are formulated in terms of causes and effects. Our intuitions are far less straightforward with respect to causally *relevant properties*. It is not obvious whether instantiations of properties can exclude each other, or overdetermine effects, in the *same way* as causes. To assume that they can is a substantial—and controversial—additional assumption. That is why I introduce the causal exclusion argument purely in *extensional* terms: formulated in terms of causes and effects only. The three basic assumptions behind the argument are the following.

(1) Mental Causation: Mental phenomena cause physical phenomena.

(2) Causal Closure of the Physical: Every physical effect has a causally sufficient physical cause.[33]

(3) Exclusion of Causal Overdetermination: Causal effects are, usually, not causally overdetermined.

The causal exclusion argument goes as follows. Every version of physicalism is committed to the three claims just presented. A non-reductive version of physicalism, however, is incompatible with the conjunction of them. Given that only events can be causes, (1) says that some mental events have physical effects. Assume that the mental event m is a cause of the physical event p; m is either a partial or a sufficient cause of p. Applying (3), we exclude that m overdetermines the occurrence of p (we assume that p has only *one* sufficient cause, if it has a sufficient cause). So, if p has a sufficient cause, c, then m either is c, or m is not a sufficient cause of p. According to (2), p has a sufficient *physical* cause. Hence, c is a sufficient physical cause of p. Given that, m cannot be a sufficient cause, but it must be a partial cause of p. Partial

[33] Generally, by *causally sufficient* I mean sufficient either for the occurrence of the effect or sufficient to determine its chance. Here, however, I have to restrict my considerations to the deterministic case. So, in what follows *causally sufficient* means sufficient for the occurrence of the effect.

causes are parts of sufficient causes. Since *c* is the only sufficient cause of *p*, *m* must be part of *c*. But since *c* is a physical—that is, non-mental—cause, *m* must be a physical cause of *p*, contrary to the assumption that *m* is a *mental* cause of *p*. The argument shows that the assumptions (2) and (3) *exclude* the causal efficacy of mental events, contrary to (1).

The exclusion argument is generally considered to be a very powerful argument that constitutes a serious problem for non-reductive physicalism. It is acknowledged that there are different versions of non-reductive physicalism that require different versions of the argument. But it is usually thought that differences with respect to the details do not diminish the main thrust of the argument. I will distinguish between three different versions of the argument, and it will emerge that the differences between them are significant.

But before that let us briefly consider a well-known response on behalf of non-reductive physicalism. That is a *reductio ad absurdum* to the conclusion that there must be *something* wrong the argument. There is no obvious reason to deny that the argument applies to the so-called special sciences in general. If the argument can be generalised, it entails that the efficacy of, for instance, chemical, physiological and biological events is excluded by the efficacy of physical events—which is absurd. That makes, as Peter Menzies has put it,

> […] a mockery of the enormous efforts devoted in the special sciences to formulating experimental and observational methodologies for testing causal hypothesis. It would follow from this position that all these efforts are misdirected because they could not, by definition, reveal anything about the nature of causal processes.[34]

The lesson to be learned is that *something* must be wrong with the argument.[35] On the basis of that one may argue that our intuitions concerning the metaphysics of causation are not reliable. If we want to settle the question whether the mental makes a causal difference, we should, therefore, not engage in further metaphysical speculation. Rather, we should consider whether it can be established that mental events have a status comparable to the status of, say, chemical, physiological, or biological events. Instead of wasting our time with the metaphysical riddle of mental causation, we should ask whether or not psychology is a special science on a par with

[34] Menzies, 2001, p. 5.
[35] See, for instance, Fodor, 1989, and Menzies, 2001.

chemistry, physiology or biology. And if it is not, we should ask whether the causal relevance of mental events can be grounded in the generalisations of common sense psychology or by virtue of the fact that they support the right counterfactuals.

Needless to say, proponents of the exclusion argument dismiss such considerations as evasive and unsatisfactory. The problem of mental causation, they insist, is a metaphysical problem that lies at the very heart of the mind-body problem—and they insist that it requires a metaphysical solution. To dismiss the problem and the attempts to solve it as metaphysical speculation, they will respond, is to brush the philosophical problem under the carpet.[36]

I agree, on the one hand, that the *reductio* shows that something is wrong with the exclusion argument. But, on the other hand, I think that the metaphysical issues cannot be dismissed so quickly. Non-reductive physicalists should at least try to spell out *what* goes wrong.

Events and Property Instantiations

I assumed that the relevant mental phenomena are mental *events*. Philosophers of mind, though, often talk about mental *states* and *properties*. It is common to use the term *mental events* in a broad sense that includes *mental states*.[37] We would, then, obtain two versions of the argument; one in terms of mental events, and the other one in terms of the instantiations of mental properties (for no one should expect the properties themselves to have a causal role). Little significance has been given to this distinction. It has been assumed that the argument is equally compelling in both versions. Let us first consider the version in terms of mental events.

The causal exclusion problem would dissolve under the assumption that mental kinds are identical with physical kinds. But this solution is obviously not an option for non-reductive physicalism, which is committed to the rejection of type-identities. Another possibility, though, is identity between mental and physical event-*tokens*.

There are two ways to distinguish events as *mental* events. An event is a mental event just in case it has a mental description, or, alternatively, just in case it has a

[36] Compare, for instance, Kim, 2000, and Crane, 1995.
[37] Compare for instance Horgan and Tye, 1988, who say that they are following a 'frequent recent practice' by using 'event' in a sense that includes 'states, process, and the like' (p. 427).

mental property.[38] I shall assume that the two definitions are equivalent. Given that, it is certainly possible that mental events are identical with physical events, even though mental properties cannot be identified with physical properties. For it is possible that one and the same event has both a mental and a physical description (or property).[39] Further, it is possible that the physical properties of such an event include those properties that realise and determine the event's mental properties. That is, it is possible that it is one and the same event that has both the mental properties and the physical properties on which they depend. Suppose that this is generally the case. Then the non-reductive physicalist can say that the mental and physical events are not rival causes, because they are one and the same event—being token identical, they share causal powers.

If we consider property instantiations instead of events, that move—the appeal to token-identity—is not available. For what goes hand and hand with that alternative version is a particular view on the nature and individuation of events, according to which events *are* property instantiations. On that view, events are properties instantiated by a substance at a time. That rules token-identity out, since the instantiation of a mental property and the instantiation of the underlying physical property constitute *distinct* event-tokens.

It seems that we have identified a significant difference between the two versions of the exclusion argument. But it merely seems so, the opponents insist. According to Davidson's theory, which is *the* token-identity theory, the events in question instantiate a causal law only under their physical description. We are warranted in regarding them as cause and effect only insofar as they are covered by physical law. But that means that they are causally efficacious only in virtue of their physical properties; they cause what they cause *not* in virtue of being *mental* events. Therefore,

---

[38] One may wonder how events can possibly have *mental* properties. It seems clear that events have or instantiate properties. The rising of Sue's arm, for example, has the property that it takes place on planet earth, that it involves Sue, and so forth. But can events have *mental* properties? Only agents, it seems, can have desires, beliefs, feelings, and so forth. Many philosophers simply assume that it is possible that events can or cannot be causally efficacious in virtue of their mental properties. I think, though, that the talk about mental properties of events is best construed as being *elliptical*. No matter what theory of events is employed, the theory in question must accommodate the fact that the events which we are concerned with are agent-involving events. *Strictly speaking*, then, a mental property of an event is a property which is instantiated by a substance involved in that event. For the sake of simplicity, though, we will talk about the properties that are had or instantiated by the event (compare Braun, 1995, p. 449, who makes a similar suggestion).

[39] Davidson has famously argued that this not only possible, but actually the case: every event that has a mental description has a physical description. See Davidson 1980, essays 11 and 12.

appeal to token-identity does not help, and the difference between the two versions is of no significance. Consider, for instance, how Stephen Yablo has put that point.

> To reply with the majority that mental events just are certain physical events, whose causal powers they therefore share, only relocates the problem from the particulars to their universal features […]. Mental events are effective, maybe, but not by way of their mental properties; any causal role that the latter might have hoped to play is occupied already by their physical rivals.[40]

A few things, however, have been overlooked in this diagnosis. In my alternative analysis I will distinguish between *three* versions of the argument, whereby each version results from a combination of non-reductive physicalism with a particular view on the individuation of events.

Version One

The first version of the argument is directed at the already mentioned token-identity theory. On that view, the mental events and the underlying physical events, which realise and determine the mental events, are token-identical. The charge against this view is that such events cause their effects only in virtue of their physical properties – the events in question have physical effects, but not *in virtue of* being *mental* events.

Let us assume, for the sake of the argument, that events cause their effects *in virtue of* some of their properties.[41] Why do the events in question have their effects not in virtue of their mental properties? How can the opponent justify the claim that they are efficacious only in virtue of their physical properties? We can distinguish between two arguments for that claim.

The first argument appeals to a connection between the causal role of properties and causal laws. It is assumed, firstly, that an event *c* causes the occurrence of the event *e in virtue of* having the property *P* if and only if *P* figures in the causal law that covers *e* and *c* (more precisely, if and only if there is a causal law according to which an event's being *P* is nomologically sufficient or relevant for the occurrence of *e*). Further, if *c* causes *e* in virtue of having *P*, *P* is said to be a causally *relevant property*

[40] Yablo, 1992, pp. 248-249.

[41] Davidson's own response is that, given his view on causation and the individuation of events, it simply does *not make any sense* to say that a cause is efficacious *in virtue of* some of its properties. Causation is an extensional relation between *events*. Certainly, in order to obtain a causal explanation we have to refer to the events using the right descriptions. But it does not follow that they are efficacious in virtue of some property (Compare Davidson, 1993, especially pp. 12-13). Opponents will either insist on the principle that causes are efficacious in virtue of some of their properties, or they will try to show that Davidson's response merely relocates the problem in a way that does not help to save the view. My task, however, is not to decide on the tenability of Davidson's response.

with respect to *e*'s causing *c*. And it is assumed, secondly, that there are no psychological laws or regularities that ground causal claims about mental events; it is assumed, in other words, that psychological *anomalism* is true. From that it follows that no event causes an effect in virtue of having a mental property.

That first argument, however, is not decisive for the following reasons. Firstly, the second assumption—psychological anomalism—is rather controversial. It is in need of independent justification and cannot just be assumed. The problem is that most philosophers who deploy the exclusion argument against non-reductive physicalism also argue *for* some form of reductive physicalism.[42] But it is hard to see what *independent* reasons a *reductive* physicalist might have to hold the second assumption. In fact, reductive physicalism seems to be committed to the claim that there are psychological laws or regularities. Secondly, the argument shifts the focus of the debate in a way that is problematic for the opponent of non-reductive physicalism. Typically, opponents of non-reductive physicalism argue that the problem of mental causation requires a *metaphysical* solution, and that all proposals that establish merely the *explanatory* relevance of mental events or properties are inadequate or beside the point. The problem, they insist, concerns causal efficacy, not causal explanatory relevance. The argument, though, shifts the focus from the causal efficacy of mental events to the question whether mental properties figure in causal laws and the question whether there are psychological laws; questions which concern the explanatory relevance of the mental, rather than their causal efficacy.[43]

According to a second line of argument, it becomes *obvious* that the events in question have their effects only in virtue of their physical properties once the principle of the causal closure of the physical is understood in the right way. According to Kim, the basic idea behind that principle is that, for every physical event, one will never 'leave the physical domain', if one traces out its complete causal history.[44] The causal history of any physical event consists only of other physical events. Events are *physical* events in virtue of having physical properties. So, the causal history of any physical event can be given in terms of other events and their

[42] The most prominent proponent of that strategy is Kim, 1997 and 2000.
[43] Further below, p. 134, I will say more about causal efficacy and causally relevant properties. Compare also note 45 below.
[44] Compare Kim, 2000, p. 40: 'One way of stating the principle of physical causal closure is this: If you pick any physical event and trace out its causal ancestry or posterity, that will never take you outside the physical domain.'

physical properties only. And that means, it seems obvious, that all the causal relations that constitute that history hold only in virtue of physical properties; there is no room left for mental properties to do any additional causal work.

The closure principle, however, is a purely metaphysical and extensional principle; it talks about causes and effects only. Assume, once more, that events and only events can be causes and effects. Consider a physical event *e* that is caused by *c*, and assume that *e* is both a physical event in virtue of having the physical property *P* and a mental event in virtue of having the mental property *M*. What licences the claim that *c* causes *e* only in virtue of having *P*?

I fail to see why and how the closure principle, by itself, rules out the possibility that *c* causes *e* in virtue of having *M*. Note, firstly, that in order to hold that *c* causes *e* in virtue of having *M* one does not have to 'leave' the physical domain. Metaphysically speaking, we cannot leave the physical domain, since each mental event is, by assumption, identical with a physical event. Secondly, the thought behind the argument cannot be that the causal relevance of mental properties is *excluded* by the causal relevance of physical properties. Without the introduction and justification of further metaphysical principles, the notion of exclusion—just like the notion of overdetermination—applies only to causes (that is, events).[45] The closure principle does not entail anything with respect to the causal relevance of properties. It does not exclude the relevance of mental properties, since it does not say that physical properties are sufficient—whatever that might mean. It says that every physical event has a sufficient physical cause. But it is an *event* that is sufficient, not an event's having a certain property rather than another.[46]

---

[45] Stephen Yablo says that although 'causes and effects are events, properties as well as events can be causally relevant or sufficient' (Yablo, 1992, p. 247, note 5). It is correct that we can talk about causally relevant events as well as causally relevant properties. But is not obvious that the sense of *causal relevance* is the same in both cases. An *event* can be causally relevant in the extensional sense of being a *partial* cause. In that sense, events are causally relevant as causes. If one denies, as Yablo does, that properties can themselves be causes, it remains to be explained in what sense properties can be causally relevant. And whatever that sense is, it is different from the one in which events are relevant, and it is, therefore, not clear at all that the exclusion argument can be restated simply by substituting *property* for *event*, as Yablo suggests.

[46] Some may construe the closure principle as saying that the occurrence of any physical event can be *explained* in physical terms only (only in terms of events and their physical properties). But by reading the argument in that way, the opponent is again shifting the focus from causation to causal explanation. The closure principle is a metaphysical principle. It is about causes, not about causal explanations. It does not say that physical causes have their effects only in virtue of their physical properties, nor does it say that everything can be causally explained in terms of physical properties.

In conclusion we can say the causal exclusion argument is not decisive against the token-identity version of non-reductive physicalism. Opponents, however, will object that even if the response is correct, it shows only that it is possible that some events have their effects in virtue of their mental properties—it shows only that mental properties can be causally *relevant*. What needs to be shown, though, is that the mental is causally *efficacious*.

What kinds of things can be causally efficacious? Causes, I suggest, and only causes are causally efficacious. And as the standard view has it, events and only events are causes. Instantiations of properties, on the other hand, can have two different causal roles. Firstly, as being the property *of* an event—as being instantiated by an event—a property can be causally *relevant* in the sense that the event causes the effect in virtue of having the property. Secondly, as *being* an event the instantiation of a property—by a substance, at a time—can be causally *efficacious*.

On the present version of the argument, mental event tokens *are* physical event tokens. That is, the instantiation of the mental property and the instantiation of the physical property do *not* constitute two *distinct* events, since they are instantiated in one and the same event. Given the distinctions just made, it follows that instantiations of mental properties can at best be causally *relevant*—they are not and cannot be *causes*. And that means that they cannot be causally efficacious. Further, instantiations of physical and mental properties cannot be causal rivals, since causal exclusion or overdetermination presupposes distinct *causes*. Now, what does *not* follow from all that is that mental *events* are not, or cannot be, causally efficacious.

Version Two

On the first version, mental and physical properties are properties of one and the same event. If we construe events as property instantiations, we obtain two further versions of the argument. Recall that that we can distinguish between properties that constitutive of events from properties that merely modify events. The former are instantiated by a substance at a time and the latter are instantiated by an event.[47] Now, in some cases it will be controversial whether a given property is constitutive of an event or not. Consider again Kim's example of *Sebastian's stroll* and *Sebastian's leisurely strolling*. Are there two events happening at the same time, or is there only

[47] Compare Kim, 1993, p. 43, and see also chapter 2, pp. 101.

one event has that more than one true description? According to what is known as the Anscombe-Davidson view on the individuation of events, there is only one event that can correctly be described as a stroll and as a leisurely stroll. According to the view that events are property instantiations, however, whether there is only one event depends on whether *strolling leisurely* is a constitutive property or whether it merely modifies the event. Recall further that we can distinguish between two ways in which the events might be related, because two events, as Kim says, can be distinct without being *entirely* distinct.[48] Sebastian's stroll and Sebastian's leisurely strolling are, according to Kim, distinct events. But they are not entirely distinct, since the latter event metaphysically *includes* the former.

Given that, we can now formulate the second version of non-reductive physicalism, according to which the instantiations of mental and physical properties constitute events that are distinct, but not entirely. Rather, physical events *include* mental events.

Kim did not try to spell out how the relation of inclusion between distinct events has to be understood; he says that the notion is intuitively plausible enough.[49] Stephen Yablo, though, has suggested the following. According to Yablo, the relation between mental and physical events is best understood as one between *determinables* and *determinates*, which, in turn, is best understood as an event's essence *subsuming* the other event's essence. What does that mean? Consider the determinate *being red* of the determinable *being coloured.* Being red is a specific—or *determinate*—way of being coloured. Similarly, Sebastian's leisurely strolling is a determinate of Sebastian's stroll, and the former subsumes the latter. The important thing to note is that the relation of subsumption—of one event's subsuming another event—is precisely the relation of inclusion. I will not go into any further detail of Yablo's account.[50] Rather, let us see how it can be applied to the problem of mental causation. The lesson to be learned, according to Yablo, is that mental and physical events cannot causally exclude each other, if they stand in the suggested metaphysical

[48] The distinction between distinct and entirely distinct events is supposed to block a counterintuitive proliferation of events. Compare ibid., pp. 42-46.
[49] Ibid., p. 45.
[50] As indicated, Yablo suggests to construe the relation of inclusion as a relation between the event's essences: an event *e* is said to subsume or include another event *e**, if the essence of *e** is a 'subset' of the essence of *e* (see Yablo, 1992, section 5, especially pp. 261-262).

relation of inclusion.[51] For we *know*, as Yablo says, that determinates and determinables are not causal rivals. True, by citing an object's being red we may be able to causally *explain* an event, or an event's having a certain property, which cannot be explained by referring to that object's being coloured. But it cannot plausibly be suggested that these properties are *rival causes*, simply because the particular instance of being red *just is* the particular instance of being coloured. If we apply this lesson to the problem of mental causation, we can then say with Yablo that 'any credible reconstruction of the [problem] must respect the truism that determinates do not contend with their determinables for causal influence.'[52]

It does not matter whether or not Yablo was successful in analysing the relation of metaphysical inclusion correctly. For, on any account of that relation, if the physical event determines the mental event by including it, the two events cannot plausibly be causal rivals. To say they cannot be causal rivals is to say that they can neither exclude each other, nor overdetermine an effect, and we can conclude that the causal exclusion argument does not apply to the second version of non-reductive physicalism.

Version Three

On the third version of the argument, instantiations of the mental and the physical properties constitute *entirely distinct* events. None of the responses that have been given so far apply. The mental and physical properties in question are constitutive of metaphysically distinct events. Instantiations of them can therefore be efficacious as causes. Furthermore, it seems that they can be causal rivals, since they constitute *entirely* distinct events.[53]

---

[51] One may wonder how a physical event can *include* a mental event. The best way to think about it is in terms of multiple realisation. Assume that systems of type *S* and type *T* can be in mental states of type *M*, and that systems of the two types realise *M* in different ways. Then the states that realise *M* in systems of type *S* and *T* must be alike in certain respects, because they both realise *M*. But given that *M* is multiply realisable, we can assume that those states also differ in some respects; they realise *M* in different ways. Given that, we can say that the way in which the two systems realise *M* are specific or *determinate* ways of being in *M*.

[52] Ibid., p. 259.

[53] One may think that this way of construing the relation between mental and physical events is not compatible with non-reductive physicalism. According to non-reductive physicalism, mental events depend on physical events. This dependence, one may think, is incompatible with the claim that the events are entirely distinct. But is it impossible that two events are entirely distinct *and* dependent? Consider the following. Brutus killed Caesar by stabbing him with a knife. At one point, Kim suggests that *Brutus' killing Caesar* and *Brutus' stabbing Caesar* are entirely distinct events, whereas *Brutus' stabbing Caesar with a knife* merely modifies *Brutus' stabbing Caesar* (compare Kim, 1993, p. 44).

Yablo argued that, if two events are not entirely distinct, in the sense that one includes the other, then they cannot be causal rivals. It does *not* follow that two events can be causal rivals, if they are entirely distinct. Nor is it obvious that all causes that are entirely distinct causes of one and the same effect are, or can be, rival causes. In order decide whether they are or can be rival causes, we need to have a closer look at some of the issues involved.

Intuitively, for causal rivalry, exclusion or overdetermination to occur, there have to be two or more *independent* causes of one and the same effect. If that is not the case, if the causes in question are not independent in the relevant sense, the exclusion argument does not apply, *because* the notions of causal rivalry, exclusion and overdetermination do not apply. In other words, whatever is necessary for causal rivalry, exclusion, and overdetermination, is necessary for an application of the argument.

In the following I will suggest a characterisation of causal overdetermination and of the involved notion of independence. I will argue that mental and physical events, construed as entirely distinct events, cannot overdetermine their effects, because they are not independent in the required sense. Given that, it is then the opponent's burden to clarify why and how the exclusion argument applies to non-reductive physicalism and to the problem of mental causation in general. Let me begin with two observations.

Let us consider a clear case of causal exclusion; a case in which two causes overdetermine an effect. An example that is often used to illustrate causal overdetermination is the case in which two sharpshooters kill their victim, which happens to be one and the same person, at exactly the same time. Each shot is sufficient to cause the victim's death, which is overdetermined by two distinct sufficient and independent causes.

The first observation is that the events in the case of mental causation are not distinct in the same way as the relevant events in the sharpshooter case. In the sharpshooter case, the events are not only distinct with respect to the instantiation of

---

Intuitively, though, the particular killing of Caesar depends on Brutus' stabbing him. To decide whether the third version is consistent would require a detailed discussion of the involved notions of metaphysical distinctness and dependence, which is beyond the scope of this thesis. Rather, I will assume, *for the sake of the argument*, that it is consistent. If it is not consistent, so much the worse for the proponents of the exclusion argument—for then the third version collapses into either the first or the second version.

properties, but they are distinct in the further sense that the properties are instantiated by distinct *substances*. This, of course, is not the case for mental causation; the mental and the physical property are instantiated by one and the same substance. That does not show, of course, that being instantiated by distinct substances is necessary or sufficient for causal rivalry. But it *suggests* that the events' being entirely distinct is not sufficient. What is required, in addition, is that they are *independent*, in some sense.

This brings us to the second observation. What we are assuming, and what the opponents of non-reductive physicalism usually assume as well, is that the two purported causes—the mental and the physical event—stand in some intimate relationship. All versions of non-reductive physicalism assume that mental events supervene, in some sense, on physical events, and that mental events are realised by them. It is assumed, in other words, that mental events (or properties) are, in some sense, dependent on physical events (or properties). *By hypothesis*, that is, the mental and their underlying physical events (or properties) are *not* independent.

This independence, however, concerns the *existence* of mental events and properties. Whenever there is a mental event, there is some physical event that realises it. It is possible that there are physical events and no mental events, but it is impossible that there are mental events and no physical events. The mental depends on the physical and is realised by it, but not *vice versa*.

What we are interested in, however, is independence of *causes* as a criterion for the application of the causal exclusion argument. Recall that the third assumption of the causal exclusion argument concerns causal overdetermination. The argument applies only if the notion of causal overdetermination applies. That is, for the argument to apply, there must be causes that *can* overdetermine an effect.

I suggest having a closer look at the notion of causal overdetermination in order to get a better grasp of the relevant notion of independence. Let us begin with a characterisation of overdetermination, which is adopted from Trenton Merrick's *Objects and Persons*.[54]

Some effect *e* is *causally overdetermined* if and only if:
(a) *e* is caused by *c*,

[54] Compare Merricks, 2001, p. 58, who thinks that a definition along such lines is 'the most literal, straightforward and natural' definition of causal overdetermination.

(b) *c* is causally irrelevant as to whether some other numerically distinct cause, *c**, is a cause of *e*, and
(c) e is caused by *c**, and *c** is numerically distinct from *c*.

Where *c* is *causally irrelevant* as to whether *c** is a cause of *e* if and only if:[55]

(d) *c* is not a cause of *c** (that is, *c* is neither a sufficient nor a partial cause of *c**, and *c* is not an intermediate in a causal chain that runs from a cause of *c** to *c**), and
(e) *c* does not cause *c** to cause *e* (nor does it cause any members of *c** to cause *e*, if *c** consists of more than one cause jointly causing *e*).

One can agree with Merricks that this extensional construal of overdetermination is straightforward. But, unfortunately, it does not help us any further for two reasons. Firstly, Merricks talks about *objects* as being causes. For overdetermination to occur, two distinct causes have to cause the effect, and distinct causes are, on that view, distinct objects. That means that rival causes—causes that can overdetermine an effect—are numerically distinct substances. That reflects my intuition that we get a clear sense of the phenomenon of overdetermination, if the causes are associated with distinct substances. But since it does not cover the case in which two distinct *events* occurring *in* one and the same substance, it is of no help.

Secondly, the definition recognises only *causal* relations between the two potentially rival causes; what matters, according to clause (b), is whether one cause is causally relevant or irrelevant as to whether the other cause brings about the effect. What we are looking for is a characterisation of the way in which the two causes of the effect, *c* and *c**, can said to be dependent or independent causes. The presented definition captures the dependence between them in terms of causal relevance; that is, the dependence is construed as *causal* dependence.

Some philosophers have tried to understand the relation between mental states and their physical realisations in causal terms. Most philosophers, however, think that the dependence between the mental and the physical is of a different kind. The notion of supervenience, as many think, provides merely a minimal constraint for that relation, which is itself in need of explanation (given that it holds). A popular solution is to say that supervenience holds, because the physical states *realise* the mental states.[56] What is needed, then, are dependence-conditions that can be applied to different relations between the physical and the mental, such as causation,

[55] Compare ibid., p. 57.
[56] Compare Kim, 2000, pp. 23-24.

supervenience, determination and realisation. An obvious candidate for that is *counterfactual* dependence.

It is commonly assumed that there is a close connection between causal dependence and counterfactual dependence. Causal relations are said to entail or support relations of counterfactual dependence. On Merricks' definition, *c* and *c** do *not* overdetermine *e*, if *c* is causally relevant as to whether *c** causes *e*. If *c* is causally relevant in that way, then whether *c** causes *e* depends causally on *c*. This causal dependence entails or supports counterfactuals of the following form:

> (CF) Given relevantly similar circumstances, had *c* not occurred, *c** would not have caused *e*.

My suggestion is to replace the causal dependence employed in Merrick's definition by counterfactual dependence. For what is crucial, it seems, is not the causal connection itself, but the entailed counterfactual. Consider again the case of the two sharpshooters. It is a case of overdetermination, because there are two sufficient and independent causes of the same effect. Specified in causal terms, the independence consists in the absence of causal connections between the two causes. Why is that relevant to the question whether the effect is overdetermined? If one shooting depends causally on the other shooting, as one might say, then the two shootings are not independent causes of the victim's death. But that is plainly circular, given that we want know in virtue of what causes are independent causes. A better answer is the following. If the two shootings are causally dependent, then one of the two sharpshooters would not have killed the victim, had the other one not done so.

The case of the two sharpshooters is a case in which it is a *coincidence* that two events cause the same effect. A causal connection would render the two shootings non-coincidental. But so would a counterfactual connection. That is why I suggest replacing the causal condition on overdetermination by a counterfactual one. Accordingly, the two causes, *c* and *c**, do *not* overdetermine the effect *e*, if *c**'s causing *e* is counterfactually dependent on *c*; that is, if CF holds.

We found Merrick's definition of causal overdetermination to be too narrow, since it covers only cases in which two potentially rival causes are causally independent. To construe the kind of independence that is necessary for causal rivalry and overdetermination in terms of counterfactual independence has the advantage that it

covers relationships of different kinds that might hold between events. In order to see whether the counterfactual approach will get us any further with our problem, let us see whether the relation of realisation—as it is usually employed by non-reductive theories—supports counterfactuals of the right sort.

Suppose that *p* is the physical realisation of the mental event *m*, and that both *p* causes *e* and *m* causes *e*. Is it true that had *p* not occurred, *m* would not have caused *e*? Let us assume that *m* is multiply realisable.[57] We have then to distinguish between two cases. In the first case we assume that each system or agent falls under a certain type or species such that for all mental event-tokens *m* of type *M*: for any agent of a certain type, tokens of *M* are realised by physical event-tokens *p* of type *P*. In that case, we can limit our considerations to agents of certain types, and restricted counterfactuals of following form will hold.

> (CF') For all agents *s* of type *S*: had the *s*-involving event *p* not occurred, the *s*-involving event *m* would not have caused *e*, given relevantly similar circumstances.

That counterfactual holds, because the occurrence of *m* depends counterfactually on *p*. If the antecedent holds, then the mental event *m* does not occur—and if *m* does not occur, then, trivially, *m* does not cause *e*.

In the second case, we do not assume that mental events of a certain type are realised by all agents of a certain type in the same way—realised by physical events of the same type. In that case we have to consider the set $\{P_1, P_2, P_3 \ldots\}$ of all possible realisations of the mental state type—across all individuals of all agent-types. In that case, counterfactuals of the following form will hold.

> (CF'') For any agent *s*: had none of the *s*-involving events p_i (of type P_i) occurred, the *s*-involving event *m* would not have caused *e*, given relevantly similar circumstances.

Again, that counterfactual holds, because the occurrence of *m* depends counterfactually on p_i. If the antecedent holds, then the mental event *m* does not occur—and if *m* does not occur, then, trivially, *m* does not cause *e*. What we get, in both cases, is a counterfactual dependence between *p*, *m*, and *m*'s causing *e*.

---

[57] Note that non-reductive physicalism is typically motivated by the possibility of multiple realisation of mental states. It is therefore safe to make that assumption. (Moreover, if relevant counterfactuals hold in case mental states are multiply realisable, they certainly hold in case they are not.)

Given all that, we can conclude that mental events and their physical realisations do *not* overdetermine their effects since they are not independent causes. The mental event is not an independent cause, because its occurrence and its causing the effect depends counterfactually on its physical realisation. No dependency of that sort holds for standard examples of causal overdetermination. The two shootings do not depend on each other—neither causally nor counterfactually. What makes the sharpshooter case a case of causal overdetermination is precisely the fact that had one sharpshooter not killed the victim, the other one would have. They kill their victim independently of each other—and that they do so at the same time, with the same success, is a mere coincidence.

I argued that the causal exclusion argument is not decisive against non-reductive physicalism, no matter whether the mental and physical events in question are construed as token-identical, distinct or entirely distinct. What is important to note is that the response to the last version applies to all three versions, since it relies on the claim that mental events are dependent on physical events—a claim that all versions of non-reductive physicalism are committed to.

What causal exclusion and rivalry amount to is fairly straightforward in case there are distinct and independent causes. But the case of mental causation is not of that sort. It is not clear at all what causal exclusion and rivalry amount to, given that the mental depends on the physical. There would have to be a kind of causal exclusion that is either not tied to the outlined notion of overdetermination, or one that is not tied to any notion of overdetermination at all. In any case, it remains to be shown what sort of causal exclusion that is and how it has to be understood. We can conclude that the opponents of non-reductive physicalism have yet to show how the causal exclusion argument applies to mental causation, even if it is assumed that mental events and their physical realisations are entirely distinct.

A First Objection

Now I will turn to two general and fundamental objections, which address the way in which the problem of mental causation and non-reductive physicalism have been presented, rather than the response to the causal exclusion argument. According to non-reductive physicalism, mental states and events depend on physical states and events, and that is why considerations of causal overdetermination do not apply. In his *Mind in a Physical World* Kim addresses that response, and he acknowledges that

the way in which non-reductive physicalism construes the relationship between mental and physical causes does not constitute a paradigmatic case of causal overdetermination. But it is not clear, Kim argues, why that removes the problem.

> The exclusion problem doesn't go away when we recognise the two purported causes as in some way related to each other, perhaps one being dependent on the other.[58]

That is because the exclusion problem, as Kim suggests, is 'not exactly that of causal overdetermination'. The problem is, rather, to show what *further* or *additional* causal work is left for a mental event, given that its physical realisation is, by itself, causally sufficient for the effect.

> [There] is a real problem, the exclusion problem, in recognising [mental] properties as causally efficacious in addition to their realizers.[59]

That is to say that the fact that the mental depends on the physical cannot be part of the solution to the problem. Far to the contrary, that dependence is the root of the problem.

> It is important to see that the [exclusion problem] arises *because* the two putative causes are *not* independent events.[60]

The thought is, it seems, that because mental events are not *independent causes*, they are not, and cannot be, *causally efficacious* in addition to their physical realisations. Hence, to point out that mental and physical events are not independent causes does not count in favour, but *against* non-reductive physicalism.

My first response is that Kim is simply begging the question. What I have offered in response to the causal exclusion argument are responses to an *argument*, which is based on assumptions that every version of non-reductive physicalism is supposed to be committed to. In that argument, considerations concerning causal overdetermination play a central role. What Kim says, however, is that the problem of causal exclusion is 'not exactly that of causal overdetermination'. But what he describes as the 'real' exclusion problem is based on an additional—and hidden—assumption; namely, the assumption that mental events have to be causally efficacious *independently of* and *in addition to* physical events. The problem with that

[58] Kim, 2000, p. 53.
[59] Ibid.
[60] Ibid.

assumption is that non-reductive physical is *not* committed to it. Non-reductive physicalism says, among other things, that mental states and events, which are realised by physical states and events, make a causal difference, and that psychology is not reducible. Presumably, the view is committed to the claim that mental states and events are causally efficacious. But, as far as I can see, it is not committed to the stronger claim that they are causally efficacious independently of and in addition to physical states and events.[61]

The reason why the causal exclusion argument appears to be a strong *argument* against non-reductive physicalism is that it is based on assumptions that everyone shares. What Kim is offering, however, is an assumption that would rule out non-reductive physicalism without further argument. In other words, Kim is not offering an argument at all.

But maybe I have misinterpreted Kim's objection. What Kim suggests, one may argue, is not that non-reductive physicalism *in particular* is committed to the assumption that mental events are causally efficacious independently of and in addition to physical event. Rather, *any* satisfying account of mental causation must show how mental events can be causally efficacious independently of and in addition to physical events. (The thought behind that would be that only independent causes are really causally efficacious—events make a *real* difference, only if their causal contribution is independent from other events that cause the same effect.)

But that cannot be correct either. Kim is one of many philosophers who acknowledge that the type-identity theory of the mind would provide a straightforward solution to the problem of mental causation. On that theory, mental events are causally efficacious, because they *are* physical events. But since they are identical with physical events they are, by definition, *not* efficacious *independently of* and *in addition to* physical events. Given that the identity theory provides a solution to the problem of mental causation, it cannot be a necessary condition that mental events must be causally efficacious independently of and in addition to physical events.

[61] Non-reductive physicalism is also not committed to the claim that mental events are causes in some *special way*. It need not abandon what Tim Crane has called the *homogeneity* of causation. Compare Crane, 1995, p. 218 and pp. 231-233.

One can, of course, deny that the identity theory solves the problem of mental causation, precisely because it does not show how mental events can be causally efficacious independently of and in addition to physical events. But that response is not open to Kim. For, as we will see, Kim's own solution does not establish that mental events are efficacious independently of and in addition to physical events. Further, it will emerge that the two cornerstones of Kim's reductive solution to the problem of mental causation are in fact compatible with non-reductive physicalism.

The first cornerstone of Kim's solution to the problem of mental causation is what he calls *physical realizationism*. On Kim's view, both realisation and reduction are construed as *restricted* to species or types of systems. In other words, whether or not something is realised by or reducible to something else is relative to different species or types of systems.

> When *P* is said to "realize" *M* in system *s*, *P* must specify a microstructural property of *s* that provides a causal mechanism for the implementation of *M* in *s*; […].[62]

Further,

> […] the idea that mental properties are realized by physical properties […] warrants reductive talk like "Having *M*, for appropriate systems, *consists in*, or *just is*, having *P*."[63]

The second cornerstone of Kim's view is the principle of *causal inheritance*:

> If *M* is instantiated on a given occasion by being realised by *P*, then the causal powers of *this instance of M* are identical with (perhaps, a subset of) the causal powers of *P*.[64]

I will not go into any further details of Kim's position. But I think it is not difficult to see why Kim thinks that the causal inheritance principle holds, given the outlined view of physical realisation. If that principle holds, Kim argues, there is no problem of causal exclusion for mental causation, because if the causal efficacy of mental events *just is* the efficacy of their physical realisers, they are, obviously, not rival causes.

That solution has a lot in common with the type-identity theory solution. It *solves* the problem simply by showing that there is no problem. Since the causal efficacy of

---

[62] Kim, 1993, p. 343, pp. 363-364, and Kim, 2000, pp. 19-23.
[63] Kim, 2000, p. 24.
[64] Kim, 1993, p. 355. Compare also Kim, 2000, p. 54.

any mental event just is the efficacy of its physical realisation, there is no problem of exclusion or overdetermination, and it follows, trivially, that mental events are causally efficacious. But just as the type-identity solution, Kim's proposal does *not* show that mental events are causally efficacious *independently of* and *in addition to* physical events.

Kim's theory is a reductive one. It is not obvious, though, whether that is essential to its solution of the problem of mental causation. What is essential, so much seems clear, is the proposed physical realizationism in conjunction with the causal inheritance principle. It may be true that the former warrants reductive *talk*, as Kim says. But it is not obvious that it presupposes or entails that psychology is reducible. Consider what Kim says about the physical states and events that realise mental states and events.

> These underlying microstates will form an explanatory basis for the higher properties and the nomic relations among them; but the realization relation itself must be distinguished from the explanatory relation [which] should not be regarded as constitutive of [the realization relation].[65]

Reduction concerns, first and foremost, theories and explanations. It can be distinguished from metaphysical issues concerning the relationship between mental and physical properties, and it can be distinguished from the issue of physical realisation. The question, however, is whether the issue of reduction cannot only distinguished from the issue of realisation, but whether it is independent of it. Is it possible that the mental properties of some system are realised by some of its physical properties *and* that psychology is not reducible?

I am not aware of any argument that shows that this is impossible. It certainly seems possible that mental events are realised by physical events even though psychology is irreducible. In the section on the coincidence problem I argued that non-reductive physicalism is compatible with a functional theory that shows how mental states are realised by physical states without reducing the former to the latter. I pointed out that in order to establish irreducibility it is sufficient to establish that the lower-level theory is not explanatory of the higher-level theory (in our case, psychology). The position that I proposed says that psychology is not reducible, and it seems clear that it is compatible with both physical realizationism and with the causal

[65] Kim, 1993, p. 344.

inheritance principle—which is just to say that non-reductive physicalism is compatible with Kim's solution to the problem of mental causation.

A Second Objection

However, the response just presented gives rise to another objection. The proposed version of non-reductive physicalism says that psychology is irreducible, that there are mental properties, and that the causal efficacy of the instantiation of a mental property just is the causal efficacy of their physical realisation. This position, the objection goes, is unstable—if not incoherent—for the following reason. A higher-level theory and the causal explanations it provides are irreducible only if the theory captures causal powers that cannot be captured by any lower-level theory. Events have their effects, as we assume, in virtue of having or instantiating certain properties. So, in order to be irreducible, the higher-level theory must capture causally relevant properties that are not and cannot be captured by any lower-level theory. In other words, if a higher-level theory does not range over irreducible properties, it is, in principle, reducible, because it does not capture irreducible causal powers.[66]

On the suggested version of non-reductive physicalism, mental events *inherit* their causal powers from the physical events that realise them. But that just means that mental events do not have irreducible causal powers. And it means that psychology does not capture irreducible mental properties. But, given that, there is no reason to think that the proposed view is a non-reductive theory. In order to avoid incoherence, either the causal inheritance principle or the claim that psychology is irreducible must be abandoned. Given that, though, non-reductive physicalism faces a dilemma: either it loses the solution to the problem of mental causation by abandoning the causal inheritance principle, or it concedes defeat by giving up the claim that psychology is not reducible.

However, that objection too is based on additional and partly hidden assumptions and principles. The problem, as I will show, is again that non-reductive physicalism is not committed to them. And again, the assumptions and principles in question rule out

[66] Kim, again, is a prominent advocate of that view. In 'The Nonreductivist's Trouble with Mental Causation' Kim asks what it is for a mental property to be 'real'. In virtue of what is it a real property? The proposed answer is inspired by what Kim calls 'Alexander's dictum': *To be real is to have causal powers* (compare, Kim, 1993, p. 348). On the basis of that he arrives at the view that mental events are irreducible only if they have *irreducible causal powers* (compare p. 350).

non-reductive physicalism directly, which means that the opponent is again begging the question.

The opponent insists that any irreducible higher-order theory must range over irreducible causal powers and properties. Given the way in which that assumption is used against non-reductive physicalism the opponent must mean the following: an irreducible mental theory must range over mental properties in virtue of which mental events are causally efficacious independently of and in addition to physical events. That must be what the opponent means, because the opponent's aim is to show that the non-reductive position is subject to the causal exclusion argument. (And in order to do so the opponent must establish that mental and physical causes are distinct and independent causes of one and the same effect.)

Given that, and given what has been said in the previous section, it is clear that the opponent is mistaken. Non-reductive physicalism is not committed to the claim that mental events are causally efficacious independently of and in addition to physical events, because it says, by definition, that mental events are dependent on physical events.

Presumably, the opponent will not be satisfied with that response. Psychological explanations, it is assumed, are causal explanations. Why are they irreducible, if they do not capture irreducible causal powers or irreducible and causally relevant properties? Let us have a closer look at some of the issues involved.

Note firstly that non-reductive physicalists and their opponents employ different criteria for reduction. Non-reductive physicalists stress that reduction concerns *theories* and *explanatory power*. Their opponents, however, talk about irreducible *properties* and *causal powers*. Non-reductive physicalists think that in order to show that a higher-order theory is not reducible to a lower-level theory, it is sufficient to show that the lower-level theory is not explanatory of it. The opponents insist that more is required; in addition, it has to be shown that the higher-level theory ranges over irreducible properties that confer irreducible causal powers onto the entities that have or instantiate those properties.

Given that, one might think that the disagreement boils down to a disagreement about the conditions for reduction. One might even think that there is no real disagreement about whether psychology is reducible or not, because the two sides deploy different conceptions of reduction. One side talks about theories and

explanations, the other side about properties and causal powers. One may think, in other words, that the opponents are talking past each other.

But I do not think that this is a correct diagnosis of the situation. What the opponents of non-reductive physicalism insist on is an *additional* criterion for a theory's being irreducible. There is some agreement about what reduction is and what the criteria for reducibility and irreducibility are. What both sides can agree on is that the criteria deployed by non-reductive physicalism are necessary for reduction. And what is contentious, it seems, is the following principle concerning irreducibility.

> (R1) A higher-level theory *T* is irreducible only if *T* ranges over properties that are causally relevant and irreducible.

However, further below I will argue that R1 is not necessarily incompatible with non-reductive physicalism. If that is correct, R1 cannot account for the disagreement. Rather, what is responsible for the conflict must be R1 in conjunction with a principle that says something like the following.

> (R2) A causally relevant property *P* is irreducible only if it confers an irreducible causal power onto the entities that have or instantiate that property (only if the entities in question have irreducible causal powers in virtue of having or instantiating *P*).

So, why can non-reductive physicalists accept R1, and what can they say against R2? Against R2, non-reductive physicalists may appeal to the status of the so-called special sciences in general. Suppose that *T* is a well-established empirical theory that is not reducible to physics, and that *T* ranges over entities that are physically realised. By application of R1 we can say that *T* ranges over causally relevant and irreducible properties, and by application of R2 we can say that the entities which have those properties have irreducible causal powers.

But what does it actually mean to have irreducible causal powers? In the present context, I can make sense of that claim only in the following way. To say that entities, which are physically realised, have *irreducible causal powers* is to say they are causally efficacious *independently of* and in *addition to* the physical entities that realise them.[67] If that is correct, then *T* is subject to the causal exclusion argument. And since *T* has been selected randomly, we can generalise and conclude that all

[67] Note that the entities in question are mental events, and that to attribute *causal powers* to *events* is somewhat inappropriate. To have a certain causal power, I take it, is to have a certain dispositional property. And it seems that only *objects* or *substances* can possess dispositional properties. Therefore, I shall interpret talk about the causal powers of events as talk about their causal efficacy.

theories of the special sciences are subject to the causal exclusion argument—which is absurd.

There are the following three obvious ways to avoid that conclusion. One may deny R1, one may deny R2, or, thirdly, one may deny that there are irreducible special sciences. I maintained—and below I will argue—that non-reductive physicalists can accept R1. Given that, there is, in the absence of further argument, no reason to reject R1. The choice, then, is between the second and the third option. The choice, in other words, is between a contentious metaphysical principle and the claim that there are no irreducible special sciences. I think there is good reason to reject the metaphysical principle R2. There is considerable agreement that the special sciences are autonomous, and that they are, therefore, not reducible.[68] The opponent may object that we might be wrong about that. Maybe we are not yet in a position to see why and how all sciences are reducible to physics. But would that count in favour of the third option? Would that count in favour of R2? I cannot see why. The fact that it might turn out to be true that all special sciences are reducible does not give us reason to believe in a *metaphysical* principle that leads, *a priori*, to the conclusion that all special sciences are reducible. The reason why we should reject R2 is, then, the following. We should reject R2, because we should reject a metaphysical principle that *entails* that all special sciences are reducible. There is reason to believe that the special sciences are not reducible, and the question of whether the special sciences are reducible or not is, at least to a great extent, an *empirical* question, and not a metaphysical one.

What is missing is an explanation of why non-reductive physicalism is not necessarily incompatible with R1. According to R1, any irreducible theory must range over irreducible properties that are causally relevant. What we need to know is in virtue of what a causally relevant property is *irreducible*. R2 provides an answer to that by appeal to irreducible causal powers. Having rejected R2, we need an alternative account. The obvious alternative is to appeal to irreducible causal regularities or laws. On such an approach, a property *P* is irreducible just in case the formulation of an irreducible higher-order theory (including the formulation of the

[68] Compare for instance Dupré, 1993; Fodor, 1974; Owens, 1989; Antony & Levine 1997 and Polger 2004.

relevant higher-level laws) must mention *P*.[69] According to that, the irreducibility of properties stems from explanatory power, rather than causal power.

However, one might still be puzzled by the suggestion that a higher-order theory, which provides causal explanations, is irreducible due to its explanatory power, even though it does not range over irreducible causal powers. I have suggested that the opponent's talk of irreducible causal powers is best understood as being about higher-level events that are causally efficacious independently of and in addition to the lower-level events that realise them. What is crucial at that point is to recall the distinction between the extensional relation of causation and the intensional relation of causal explanation. Given that distinction, it is certainly possible that there are properties, which are causally relevant and irreducible in virtue of the explanatory insight provided by theories that range over them, rather than in virtue of conferring irreducible causal powers onto the entities that have or instantiate them. Consider the following.[70]

Assume that there are two mental events *m* and *m** such that we can explain the occurrence of *m** and *m**'s having the higher-level property *A* by reference to *m*'s occurrence and *m*'s having the higher-level property *R*. Assume further that *m* is realised by the physical events $p_1,\ldots, p_j$; that *m** is realised by the physical events $p_1^*,\ldots, p_k^*$; and assume that the causal process between *m* and *m** consist of nothing but causal chains leading from the p_i's to the p_i^*'s.

Given that, one is citing the causes of the p_i^*'s, if one is citing all the p_i's, and one is thereby citing the causes of the physical events that realise *m*. From that it does *not* follow that the particular instantiation of *A*—that is, the occurrence of *m**'s being *A*—can be causally *explained* by reference to the p_i's (or by reference to the complex physical event that is constituted by the p_i's). That shows that it is possible that *m*'s being *R* can causally explain something, namely an instantiation of *A*, that cannot be explained by reference to *m*'s physical realisation.

---

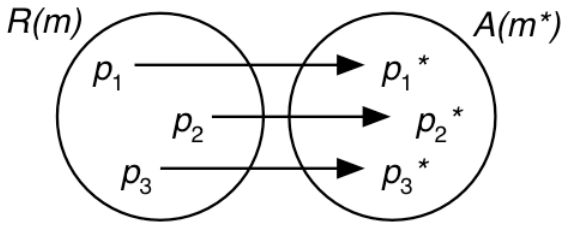
[69] Compare for instance Antony and Levine, 1997. They say that a property is 'real (or autonomous) just in case it is essentially invoked in the characterisation of a regularity' (p. 91). And they argue that it does not follow that 'real' properties have 'distinct causal powers' (compare p. 92).

[70] The following is inspired by David Owens, 1989. Owens argues that higher-level properties and regularities *cross-classify* lower-level entities. The complex lower-level physical events that realize the higher-level events are not explanatory, because from the perspective of the physicist they are like 'shapeless fusions of physical events' (p. 67).

The opponent, I suppose, will object that the mental events and properties in that scenario appear to be *merely* explanatory. All the real causal work, it seems, is done by the physical events, and the mental events do not really make a difference.

However, it is a mistake to think that the mental events are *merely* causally explanatory. Partly responsible for the underlying confusion, I think, is a misconstruction of the model of levels of explanation. The levels of explanation, of course, must be understood as different levels of description and abstraction, rather than, literally, as different layers of the world. Let us have a look at the kind of figure that is typically used to illustrate the problem of mental causation, which reflects that misconstruction. In the following figure, p and p^* are physical events and m and m^* are the mental events that are realised by them. The vertical lines stand for the relation of realisation and the arrow stands for causation.

The mental events belong to the higher-level of mental explanation L_M, and unless we add another arrow from either m to p^* or from m to m^*, the mental event m appears to be causally impotent. However, the account that I have presented looks rather different. Consider the following figure that represents the scenario discussed above.



In that figure the circles capture the physical events, which realise the mental events and the instantiations of mental properties. The levels of explanation are not pulled apart, and the mental event m does not appear as causally impotent. The particular occurrence of m and m's being R is realised by the particular p_i's, and the causal efficacy that is exercised by the p_i's is, *ipso facto*, exercised by m. The most important thing to note is that the cause is the whole complex event on the left hand side, and that the effect is the whole complex event on the right hand side. The cause is the p_i's realising m's being R. It would be a mistake to say that m is merely explanatory, or that the p_i's alone do the causal work. Again, the cause is the p_i's realising m's being

R; hence, *m*'s being *R* causally explains and *causes m**'s being *A*.[71] But even though the effect too is the whole complex event, it is possible that certain features of the effect can only be explained by referring to certain features of the cause. It is, therefore, possible that the explanatory insight gained by reference to the cause as a mental event is not reducible; it is possible that *m*'s being *R* can causally explain something—an instantiation of *A*—that cannot be explained by reference to *m*'s physical realisation.[72]

Conclusion

My aim was neither to present a fully developed and detailed account of causalism, nor to provide a full defence of causalism construed as non-reductive physicalism. Rather, my aim was to contribute to a defence of causalism by showing that two main obstacles—the coincidence problem and the causal exclusion argument—can be overcome, no matter whether psychology is reducible or not. Both the suggested solution to the coincidence problem and the response to the exclusion argument raise further questions that are beyond the scope of this work. What has been shown, though, is that with respect to both problems good suggestions and responses are available to non-reductive physicalism.

What is important to note is that the presented considerations are not only in defence of non-reductive physicalism, but that they are part of the overall argument *for* causalism. I argued that the main advantages of causalism are, firstly, that it provides an answer to the question how reasons influence or affect actions. And, secondly, it gives an integrated account of agency by providing an account of the relationship between reasons for actions and the causes of bodily movements. The

[71] According to non-reductive physicalism, the mental depends on and is determined by the physical. The relation between the mental and the physical is usually specified more precisely in terms of supervenience (compare chapter 2, p. 106, note 98). A plausible definition of supervenience should capture that any agent, whose having a certain mental property is realised by having certain physical properties, has that mental property *necessarily*; where the modality in question is metaphysical necessity. Given that, it is not possible that the p_i's cause the p_i*'s without realising *m*'s being *R* and *m**'s being *A*, respectively. For more on that see, for instance, Noordhof, 1999.

[72] I should mention an alternative strategy available to non-reductive physicalists. I argued that non-reductive physicalism is compatible with R1 (the claim that irreducibility of a theory requires that it ranges over irreducible properties). Alternatively, non-reductive physicalists might reject metaphysical speculation concerning the causal relevance and reducibility of properties altogether. In connection with that, they might also deny that there are mental *properties*. Rather, the irreducibility of the mental concerns only theories and their vocabulary. What is irreducible are not mental properties, but mental concepts, predicates and, generally, sentences featuring intentional vocabulary.

important point is that the suggested solution to the coincidence problem is not just compatible with causalism, but it *presupposes* causalism. The central idea was that a solution to the coincidence problem must show how the mechanisms described by the lower-level theory preserve the causal architecture of the higher-level theory. That, of course, presupposes that the higher-level theory describes a causal architecture. It presupposes, in other words, that reason-explanations are causal explanations. Given the absence of an alternative account of the relationship between reason-explanations of actions and causal explanations of movements, we have not only defended causalism, but we have provided reason to endorse it.

Chapter Four: The Standard-Causal Model and its Limits

In the first chapter I defended the reductive standard-causal account of agency against non-reductive alternatives. Proponents of the non-reductive approach think that the standard-causal theory cannot account for certain important aspects of human agency—namely, that we can choose to act in the light of reasons, that we can do so with free will, and that we are morally responsible for some of our actions. Against that I argued that a non-reductive theory does not do better in accounting for any of those abilities or aspects. The non-reductive theory, I argued, not only fails to account for the required kind of control, but it fails to account for an agent's exercise of control altogether. In this chapter I will return to the notion of agential control and I will respond to the more fundamental challenge that the reductive model cannot account for agency at all. I will acknowledge the challenge, but I will deny that it constitutes a genuine problem for the standard-causal theory. Provided that the theory can account for an agent's exercising control, as I will suggest, there is no reason to think that the theory cannot capture the phenomenon of agency.

After that, I will address further issues and problems concerning the standard-causal model, such as the problem of deviant causal chains, the question whether all actions that are done for reasons are based on deliberation, and the question whether the standard-causal model can account for *all* aspects of human agency. We will see that the standard-causal model cannot account for the ability to act with *libertarian* free will. I will argue, though, that we should nevertheless endorse the standard-causal model of agency, and abandon instead the belief in our having free will.

The Challenge of Disappearing Agency

Some philosophers think that the reductive standard-causal approach fails to capture the phenomenon of agency altogether. In his *Free Action*, A. I. Melden considers two versions of the view that actions have causal antecedents. On the first version, actions are bodily movements that have, as Melden says, a 'complete causal explanation […]

in terms of brain states, stimuli, muscle movements', and so forth.[1] Melden assumes that causal explanations are grounded in strict causal laws, and that causes necessitate—or causally determine—their effects. On the basis of that he argues that a causal theory of action would not only render free and morally responsible action impossible, since it entails that one could never have done otherwise. But, more fundamentally, such a view would leave, as Melden says,

> […] no room for personal agency, [for] there is nothing in this account that is 'my doing' – I am a helpless victim of the conditions in my body and its immediate physical environment.[2]

Then he turns to the view that the causal antecedents of actions are *mental* states and events. According Melden, the fact that an action is caused by mental antecedents, such as 'volitions, desires, interests etc.', does not help to render the causal theory of action any more plausible. For even on that version of the causal theory, the agent is no more than a *victim* of causal forces.

> [What] I willed [was] not something *I* really willed and did, but something that was made happen by my antecedent conditions, my mental condition, my inclinations, my desires, motives, and so on. If *these* are the causal factors and if these are subject to causal explanation in terms of antecedent psychological factors, then whatever happens is none of my doing […].[3]

What Melden expresses here, it seems, is an intuition that we encountered in the first chapter; namely, the intuition that actions are done or performed by the *agent himself*, rather than being caused by agent-involving states or events. Melden says that the 'self' in any instance of self-governance cannot be identified with, or reduced to, any arrangements of the agent's mental states and events: 'I am not any one of these [psychological antecedent] factors, nor all of them', and whatever is caused by them is 'not strictly speaking *my* doing', because the behaviour they cause 'can be explained not by reference to my *self*, but to various events in the psychological mechanism'.[4]

Melden's challenge goes beyond the objections that were considered in the first chapter insofar as it addresses agency *as such*, rather than aspects of agency—such as acting for reasons and acting with free will. Further, Melden emphasises the point that

[1] Melden, 1961, p. 7.
[2] Ibid.
[3] Ibid., p. 8.
[4] Ibid., pp. 8.

both bodily movements and their mental antecedents, which are supposed to constitute actions, are mere *happenings*. They are mere events that happen *to* or *in* the agent, rather than things that are being done by the agent. If the causal theory is correct, as Melden thinks, then

> […] each of us, myself included, as I survey the natural history of our [actions], is a victim, witting or not, of these goings-on that make all the difference to what, in our common and confused or downright mistaken way, we describe as the things that people do.[5]

According to Melden, the very fact that the standard-causal model construes agency in terms of *events*—and event-causal processes—renders it inadequate. An agent's performing an action is essentially a *doing*. It cannot be identified with, reduced to, or explained by *happenings*—that is, events—or causal relations between them. Melden does not deny that actions can *appear* as happenings. They can be described as events, if we observe their 'natural history'. But even if actions are—or appear as—events from a certain perspective, in order to understand their nature *as* actions we must assume a stance that recognises them as *doings*.

Thomas Nagel's critique of the standard-causal model focuses on that last point. Nagel thinks that actions *disappear* if we view them from the same standpoint that we take towards phenomena of the 'natural world'.

> Something peculiar happens when we view action from an objective or external standpoint. […] Actions seem no longer assignable to the individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part. […] The essential source of the problem is a view of persons and their actions as part of the order of nature, causally determined or not. That conception, if pressed, leads to the feeling that we are not agents at all, that we are helpless and not responsible for what we do.[6]

Nagel makes clear that the problem is not causal necessitation—or causal determination. Rather, the problem is that the standard-causal theory assumes an objective and external standpoint, which construes action as being part of the natural order. Let us call that standpoint, which seeks to explain all concrete phenomena in event-causal terms, *naturalism*.[7]

---

[5] Ibid.
[6] Nagel, 1986, p. 110.
[7] Arguably, that is only a minimal condition for what is otherwise known as naturalism. Naturalism, typically, makes stronger claims in addition to that (compare, for instance, MacDonald, 1992, and Pettit, 1992). However, in the context of action theory it is common to characterise naturalism in that minimal way. Compare Velleman, 2000, p. 130 and Bratman, 2001, p. 312.

Just like Melden, Nagel thinks that it does not matter whether the causal antecedents of action are mental or, say, neuro-physiological in kind, because the problem is that actions are construed as *events*, caused by other events.

> […] *my doing* of an act […] seems to disappear when we think of the world objectively. There seems no room for agency in a world of neural impulses, chemical reactions, and bone and muscle movements. Even if we add sensations, perceptions, and feelings we don't get action, or doing—there is only what happens.[8]

The challenge presented by Melden and Nagel amounts to the following. The naturalistic standpoint recognises only events and things that are associated with event-causation and event-causal explanations.[9] That standpoint does not and cannot capture agency—actions, doings or activity. Understanding itself as being part of naturalism, the standard-causal theory assumes that standpoint. Therefore, it cannot capture the phenomenon of agency.[10] Let us call that the challenge of disappearing agency.[11]

Is there a Problem of Disappearing Agency?

The first question that we have to ask is whether that challenge constitutes a genuine *problem* for the standard-causal theory. At the heart of Melden's and Nagel's challenge is the claim that naturalistic theories cannot capture the phenomenon of action or agency, because they recognise only *happenings*—that is, events. As far as that general point is concerned, Melden and Nagel do not offer any *argument*. The force of the challenge is merely intuitive, and from that alone it is not obvious that they have identified a genuine problem for the standard-causal model of agency.

Proponents of the standard-causal approach may dismiss the challenge as question-begging—or, simply, as beside the point. The standard-causal theory offers an *account* of agency in the sense that it provides necessary and sufficient conditions

---

[8] Nagel, 1986, p. 111.

[9] Things such as states or standing conditions, dispositions and the objects that possess them, facts and states of affairs.

[10] Compare also with what Bishop, 1989, calls the problem of natural agency: 'the problem of natural agency is an ontological problem—a problem about whether the existence of actions can be admitted within a natural scientific ontology' (p.40).

[11] Mele, 2003, pp. 215 and Lowe, 2003b, section 1, talk about 'the problem of disappearing agents'. Lowe refers to Gideon Yaffe, 2000, who calls it the 'Where's the Agent Problem.' The challenges of disappearing agents and disappearing agency are very closely related, since agents are beings who exercise agency. I will turn to a related challenge, which I will call the challenge of disappearing *agents*, further below.

for action—for an exercise of agency—to take place. If these conditions referred to actions, doings or related agential phenomena such as an agent's exercising control, the account would be circular. Melden and Nagel dismiss the approach because it mentions only events, causal relations and causal explanations, but no *doings*. But, of course, necessary and sufficient conditions for action must not mention doings or any other agential phenomena. What Melden and Nagel overlook is that the naturalistic approach does not claim to provide an *analysis* of our *concept* of agency in standard-causal terms, but conditions for the realisation of agency by event-causal processes.[12] The challenge is therefore off target, and certainly it does not constitute a problem for the standard-causal theory.

However, I shall not draw upon that response, since it is based on a rather uncharitable reading of Melden's and Nagel's point. That reading makes a mockery of the challenge of disappearing agency by reducing it to the claim that no event—or event-causal processes—can constitute an action, because happenings are not doings. Instead, I suggest a more charitable interpretation of the challenge in terms of *control*. Whenever an action is performed, the agent exercises some kind of control. According to Melden and Nagel, the naturalistic stance does not recognise agency. From that perspective, agents appear as a mere locus where states and events take place; they appear as beings that are subject *to* happenings, rather than as subjects *of* control. On that reading, the challenge for the naturalistic approach is to show how an agent's exercise of control can be accounted for in standard-causal terms. The challenge is, in other words, to explain why and how a particular causal process—consisting of agent-involving states and events—constitutes agential control. Note that this is merely a challenge, but not a problem for the naturalistic approach. Melden and Nagel have not presented an argument to the conclusion that a standard-causal account of control is untenable or incoherent. Rather, they have presented intuitions, which show that it is *difficult to see* why and how a standard-causal process can constitute control.

The appropriate response to the challenge, I propose, is to show that the standard-causal model has resources to *distinguish* between causal processes that constitute control and ones that do not. If that distinction can be drawn in standard-causal terms

[12] Bishop, 1989, makes that point in his defence of the event-causal theory of action; compare especially pp. 177-180.

only, then the standard-causal theory can, in effect, account for control. Opponents might complain that this response is unsatisfactory, because it will refer only to happenings. By making that point, though, they would commit themselves to the uncharitable—and question-begging—reading of the challenge, which they should dismiss rather than endorse. Let us turn, then, to the standard-causal account of agential control.

Agential Control

It seems obvious that an agent who acts *for* reasons exercises agential control. According to the standard-causal model of agency, acting for reasons requires that the action is caused and rationalised by some of the agent's mental states and events—it requires, in other words, that the action is caused by *reason-states*. However, rationalisation and *mere* causation is not sufficient. Reason-states must cause actions in the *right* way, which is highlighted by examples involving so-called deviant—or wayward—causal chains. The possibility of deviant causal chains constitutes a serious and difficult problem for the standard-causal theory of action. In the next section I will argue that this problem can be solved. The solution to that problem completes the standard-causal account of acting for reasons, and it provides, thereby, a standard-causal account of agential control. However, the solution that I shall propose will not only complete the account of acting for reasons, but it will also help us to see why that account provides us with an account of *control* in event-causal terms. It will therefore be useful to anticipate some aspects of the problem and its solution.

Typically, an argument from deviant causal chains against the standard-causal approach is presented in forms of examples in which all the standard-causal conditions are satisfied, but in which it seems obvious that the agent does not perform an action at all (or in which it seems obvious that the agent does not perform the action intentionally). Such examples are therefore presented as counterexamples to the standard-causal model of action (or intentional action). Consider, for instance, Davidson's much-discussed climber example, which introduces the most basic—and most troublesome—type of causal deviance.

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the

> rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold […].[13]

The climber's belief and desire cause and rationalise his loosening the hold on the rope. But, *obviously*, the climber does not perform an action at all, let alone the action of loosening his hold on the rope *intentionally* or *for* reasons. Hence, causation and rationalisation by reason-states is not sufficient for agency.

It is generally agreed that the standard-causal theory must exclude such cases—it must require that reason-states cause movements in the *right* or *non-deviant* way.[14] In the example, the reason-states cause the bodily movement, but not directly. Rather, there is a causal chain from the reason-states to the movement that runs through a state of nervousness, which renders it false that the agent performs an action. It is that causal intermediary—that state of nervousness, which links the reason-states and the movement—that renders the causal pathway *deviant*. The agent, it seems, fails to perform an action, because the agent's reason-states cause the movement *via* that intermediary state.

According to the solution that I endorse, cases of causal deviance can be excluded by requiring that the bodily movement is *guided* by and *responsive* to the relevant reason-states—rather than being merely caused by them. I will show, firstly, that the notions of guidance and responsiveness are compatible with the reductive standard-causal approach, and secondly, that requiring guidance and reason-responsiveness solves the problem of causal deviance. In cases like the climber example, for instance, the bodily movement is not a response to the climber's reason-states, but to his nervousness. By requiring reason-responsiveness the standard-causal model can accommodate examples of that type.

What emerges is a plausible and informative account of agential control in terms of non-deviant causation by agent-involving reason-states.[15] There are two things to

---

[13] Davidson, 1980, p. 79.

[14] Goldman, 1970, argued that the problem of characterising non-deviant causal pathways is an empirical problem. It is, though, now agreed among virtually all proponents of the standard-causal model that the problem of deviant causal chain is a philosophical problem (Goldman acknowledged that it is a problem—he merely denied that it is one for philosophy). Proponents of the standard-causal model who acknowledge the philosophical problem include Bishop, 1989, chapters 4 and 5; Davidson, 1980, pp. 232-233; Mele, 2003, 51-63; Brand, 1984, chapter 1; Enç, 2003, chapter 4; Peacocke, 1979; Searle 1983, chapters 3 and 4; Thalberg 1984.

[15] It is the agent *himself* who is in control insofar as the movements issue from his *own* reason-states—insofar they issue from what the agent *himself* desires, believes, values, intends, and so forth. There may be problems with mental attitudes that the agent has not acquired in a normal or appropriate

note. Firstly, given the outlined model of acting for reasons, we get a clear sense as to why that provides an account of *control*. According to the standard-causal model, an action is done for reasons only if it is caused, guided by and responsive to some of the agent's reason-states. We can understand why an agent whose action is guided by and responsive to what she desires, believes and intends exercises agential control by performing that action. Secondly, it seems plausible to say that states such as nervousness and agitation render causal pathways deviant, because they undermine the agent's control. The climber, for instance, is not performing an action, because his loosening the hold on the rope is triggered by something—his nervousness—over which the agent has no or only insufficient control, and the causal pathway is deviant, because it runs through that control-undermining state. A solution to the problem of causal deviance excludes such cases, and it thereby excludes control-undermining states. Given the suggested solution, then, guidance by and responsiveness to reason-states excludes control-undermining states.

Given all that, we can conclude that the standard-causal model has the resources to distinguish between an agent's having control and an agent's lacking control, and that it can, therefore, account for an agent's exercise of control. The standard-causal model can explain what control consists in and how it is realised by event-causal processes only.[16] That shows, then, that agency does not *disappear* from the naturalistic stance of the standard-causal model, since agential control does not disappear from that stance.

Acting, and Acting for Good Reasons

Opponents may find objectionable that the offered account of control is restricted to acting for reasons. Why should one think that an agent is exercising control only when he is acting for reasons? The offered account is implausible, because it is overly

way (but through brainwashing, for instance). It may be that additional conditions on *ownership* (conditions that specify in virtue of what a given attitude is the agent's *own* attitude) are necessary for moral responsibility and autonomous action. But for an exercise of a basic kind of control it seems sufficient that the mental states are the agent's own in the plain sense that the agent is the subject of those attitudes—or that the agent instantiates them.

[16] The proposed response to the challenge of disappearing agency bears similarities to a well-known compatibilist strategy. Incompatibilists about free will hold that the existence of free will stands and falls with the truth of causal determinism: if causal determinism is true, we do not have free will, and if we have free will, causal determinism is false. A standard compatibilist strategy is to point out that their opponents ignore important distinctions that can be made with respect to different kinds of causal connections. In particular, we can distinguish between deterministic causal processes that undermine freedom (processes that constitute coercion or compulsion, for instance), and ones that do not.

rationalistic. We are capable of performing actions *voluntarily* and *spontaneously* for no reason whatsoever. Examples are actions such as scratching the back of one's head, drumming with one's fingers or crossing one's legs.[17] Acting spontaneously, agents exercise a kind of control that cannot be construed as guidance by and responsiveness to reasons. Even if the proposal provides a correct account of acting for reasons, it captures only rational agency, but not agency understood as the capacity to act spontaneously.

In response to this objection I will offer a list of clarifications concerning the standard-causal model of action. In the first chapter I characterised action as activity that is intentional and rational in the sense that it is subject to rationalising explanations. It is intentional in the sense that rationalising explanations describe it as purposeful and goal-directed; they show that the agent intended to achieve some goal or satisfy some desire by performing the act.[18] It is rational in the sense that it appears as intelligible in the light of some of the agent's mental attitudes, which the rationalising explanation appeals to. I called mental states and events that help to rationalise the performance of some action reason-states, and I assumed that activity, which is subject to true rationalising explanation, is action that is done for reasons. Further, I pointed out that this characterisation goes hand in hand with the definition of action in terms of its causal history, according to which an agent-involving event is an action if and only if it is caused in the right way by agent-involving reason-states. As it stands, then, the standard-causal model identifies action and agential control with intentional action and intentional action with action that is done for reasons.

A first clarification concerns the notion of reasons for action and the corresponding notion of acting for reasons. I pointed out that the employed notions of

[17] Compare for instance Ginet, 1990: 'Many a time, for example, I have voluntarily crossed my legs for no particular reason. No antecedent motive, no desire or purpose I expected thereby to serve, prompted me to do it' (p. 3).

[18] To say that acting is intentional activity is, of course, not to say that all actions are intended or done intentionally. By raising her arm, say, Sue scares away a fly. Sue may or may not have the intention to scare away the fly. If not, it may still be true that Sue scares away the fly intentionally—say, in virtue of expecting it as a side-effect. But it seems uncontroversial that scaring away the fly is something that Sue *does*. Assume, then, that this *action* is neither intended nor done intentionally. That can be accommodated in two ways. Assuming a coarse-grained theory of events, the action is intended and done intentionally only under the description of being a raising of an arm. But scaring away the fly is caused by her intention to raise her arm as the two acts are token-identical. Assuming a fine-grained theory, Sue's scaring away the fly is an action as it is generated by the raising of the arm, but it is not intended and not done intentionally as it is not represented—as goal or as expected side-effect—in the content of the intention that causes the arm's movement. Compare Mele, 1997, pp. 233-234. For more on intentional action see, for instance, Bratman, 1987, Mele and Moser, 1997 and Enç, 2003.

reason and *rationality* are subjective or internal. What appears as rational in the light of some of the agent's attitudes may not be something that we would normally call rational; an action that appears as rational in that internal and agent-relative sense may well be irrational or unintelligible in an objective and agent-neutral sense. Further, an action that appears as intelligible in the light of some of the agent's reason-states may be irrational in the light of *all* of the agent's reason-states, as in cases of weak-willed action. The same holds for the corresponding notion of reasons for action, because acting for reasons is construed as action that is caused by reason-states in a non-deviant way. That is, an agent may act for reasons in that internal and agent-neutral sense even though there was no reason to perform the action (in the circumstances) in an objective and agent-neutral sense of what there is reason to do.

Consider, for instance, Bernard Williams's gin and tonic example.[19] Sam wants a gin and tonic and he believes that the glass in front of him contains some. In fact, though, it's not a gin and tonic, but a glass of petrol. Williams asks whether Sam has reason to drink the stuff, and he notes that 'there are two ways here' to answer that question. On the one hand, it is clear that Sam does not have reason to drink it—clearly, he does not have reason to drink petrol. On the other hand, if he drinks it, we can explain his action in terms of his desire and belief, and that explanation seems to be a reason-explanation of his drinking the petrol. Now, I do not want to claim that Sam has *normative* reason, nor do I want to deny that. All I am saying is that if Sam drinks the stuff in the glass in front of him, then he acted for a reason in the agent-relative sense, even though it may be wrong that he had normative reason to drink it—reason in the objective or agent-neutral sense.

According to the standard-causal model, then, acting for reasons is not necessarily acting for *good* or *normative* reasons. Some philosophers will protest against that claim. They will insists that all reasons for action are good reasons for action, because something is a reason only if counts in favour of something—only if it points towards something that is good about something.[20] I think, though, that this attack is beside the point. As the gin and tonic example shows, some actions can be rationalised even though there was no reason to perform that action. To say that such actions are done for reasons is a *plausible* terminological stipulation, rather than a

[19] Compare, Williams, 1981, p. 102.
[20] For instance, Dancy, 2000.

substantial claim about normative reasons. Given that, it is entirely unproblematic to distinguish between acting for reasons and acting for good or normative reasons.

The point of the first clarification is, then, that the claim that all acting is acting for reasons does not entail that all acting is acting for good or normative reasons. It means, rather, that all actions are caused in the right way by reason-states. Similarly, to construe agential control in terms of acting for reasons is not to construe it in terms of acting for good reasons.

The second clarification concerns reason-states—mental states and event that are appealed to in rationalising explanations of actions. Proponents of the standard-causal theory do not agree fully on which mental states and events are the ones that cause and rationalise actions. The popular candidates are desires, beliefs, intentions, and events of desire-, belief-, and intention-formation. The currently prevailing view is that the proximate causal antecedent of every action is the formation of an intention.[21]

Often the formation of an intention is itself caused by other reason-states, such as desires, beliefs or other intentions.[22] In some cases the formation of an intention to act may issue from a desire alone. Those are typically cases in which the agent has an *intrinsic* desire to do something—that is, roughly, a desire to do something just for sake of doing it or just because one feels like doing it. And in some cases, the formation of an intention may not be caused by another reason-state at all, but may issue directly from other mental events, such as perceptions. The sight of a chocolate cake, as Mele suggests, may directly issue in the formation of an intention to buy one, without any specific desire to do so and without any belief that buying a chocolate cake serves some further purpose.[23]

---

[21] Compare Mele and Moser, 1997. David Velleman, for instance, describes 'the standard story of human action' as follows: '[the agent's] desire for an end, and his belief in the action as a means, justify taking the action, and they jointly cause an intention to take it, which in turn causes the corresponding movements of the agent's body' (Velleman, 2000, p. 122). For detailed accounts and versions of the view see Brand, 1984; Bratman, 1987; Mele 1992 and 2003; and Enç, 2003.

[22] The claim that every action is caused by an intention seems to lead to a regress, if the acquisition of an intention is itself an action—namely, the mental act of making a decision. For then the *act* of forming an intention must as well be caused by the formation of another intention, which is, again, a mental act. However, Mele, 2003, argues that we can distinguish between intentions that are acquired *passively* and ones that are acquired *actively* by making a decision. Further, the claim is not that all actions of, for instance, type *A* are caused by an intention to *A*. It says, rather, that all actions are caused by *some* intention—it does not need to be an intention *to A*. Given that, proponents of the standard-causal model can propose that every *active* formation of an intention—every decision—is partly caused by a *passively* acquired intention to settle the practical question what to do (pp. 202-205).

[23] Compare Mele 2003, p. 201.

Given all that, it is clear that states and events of all three types—desires, beliefs, and intention—can be reason-states. It is very controversial whether beliefs alone can cause and rationalise action. But some desires and some intentions, it seems, can by themselves rationalise actions. In particular, the performance of some actions may be rationalised by referring to intrinsic desires or intrinsic intentions only. That highlights another respect in which the employed notions of reasons and rationality are minimal and agent-relative. Consider Sue's drumming a rhythm with her fingers. Given that this is motivated by an intrinsic desire to drum a rhythm, the action can be rationalised by reference to that desire and it is done for that reason. The performance of the action makes sense in the light of the desire, and it is done for that reason given that it is by motivated it. Reason-explanations of actions of that kind seem banal, superfluous, or ridiculous. But that, of course, is irrelevant to the question whether or not such actions admit of reason-explanations. Explanations of such actions strike us as banal or ridiculous, because they rationalise utterly insignificant movements, which we virtually never have to explain or justify in everyday discourse. Nevertheless, it may well be true that I scratched my head or crossed my legs, *because* I wanted to relieve a minor discomfort and because I believed that performing those actions are adequate means to that end.

Finally, the standard-causal model does not require that the agent *attends* to performing the action for a reason. It does not require that the agent consciously or reflectively deliberates about the action and the reasons in favour of it. The reason-states may not be at the forefront of the agent's mind, as it were, or they may even be operative at a sub-conscious level. And it does not require that agents act for *obvious* or *transparent* reasons—the reasons the agent acted for need not be transparent to the agent himself, nor to others.

The point of the second clarification is, then, that spontaneous actions such as scratching the back of one's head, drumming with one's fingers or crossing one's legs may well be done for reasons in the sense outlined. They may not be done for good or normative reasons and they may not be done for any further reasons, in the sense that they do not serve any further purpose. But that does not show that they are not done for reasons in the subjective and agent-relative sense of being caused by reason-states. It is, therefore, far from obvious that instances of spontaneous and merely voluntary

actions exhibit a kind of control that cannot be captured as non-deviant causation—guidance by and responsiveness to—reason-states.

Furthermore, even if there were voluntary and spontaneous actions, which are not done for reasons in the sense explained and which involve a different kind of control, it would not follow that the standard-causal model cannot capture the phenomenon of agency, because it would not follow that it cannot capture agential control. Acting for reasons is a kind of agency and agents who act for reasons exercise agential control. Given that the standard-causal model can capture acting for reason, it can capture, at the very least, a *kind* of agency and a *kind* of agential control. We can conclude, then, that is has not been shown that there is a *problem* of disappearing agency for the standard-causal theory, and that the challenge of disappearing agency can be met by the standard-causal account of agential control construed as acting for reasons.

Agency *As Such*

The opponent may object to that last point that the challenge of disappearing agency concerns the phenomenon of agency as such. If one interprets the challenge in terms of agential control, then the offered notion of control must capture the phenomenon agency as such, rather than some particular kind of agency.

My response to that is that there is no plausible interpretation of what agency *as such* is. Agency, rather, is best understood as a *family* of different kinds of agency—such as rational, deliberative or autonomous agency. In order to see that let us consider whether we can make sense of the notion of agency *as such*.

Assume that we can reasonably well distinguish between human agency and other animal behaviour. Agency *as such* could then be understood either as human action in general, or as including animal behaviour in general—including human action. The latter construal can be ruled out for two reasons. Firstly, most philosophers, including Melden and Nagel, are primarily concerned with human agency; usually they formulate their points using first- or third-person statements—referring to themselves, their readers or to human agents in general. More importantly, though, if we construe agency in that broad sense, the challenge becomes very implausible. For then the proponents of the challenge are committed to the claim that the behaviour of other animals—cats, flies, spiders or whatever you like—cannot be captured in naturalistic terms, which is absurd. The first construal also faces problems. If we construe agency *as such* as *human* agency in general, we

are restricting the phenomenon to one kind of agency, which contradicts the spirit of the objection. Far more importantly, though, we do not have a characterisation of human agency *as such*. All attempts to distinguish between human and other animal behaviour refer to some kind or aspect of agency, which is deemed distinctively human—such proposals aim to identify a specific capacity or ability that distinguishes us from other animals.

Agency, I suggest, is best broken down into kinds or aspects of agency. Accordingly, questions and problems concerning agency are best broken down into questions and problems concerning certain kinds of agency—such as purposive behaviour, acting for reasons, deliberative action, reflective and self-controlled action, autonomous and free action. This list of different kinds of agency, however, is more than just a list; it is a *hierarchy* of kinds of agency. It begins with a very basic form of agency—namely, purposive behaviour—and continues with more and more sophisticated forms—rational action, deliberative action, and so on. It is a hierarchy in the sense that an agent's having—or being capable of—each of the subsequent forms of agency presupposes that the agent is capable of the preceding, more fundamental, forms of agency. For instance, being capable of acting for reasons presupposes that one is capable of purposive behaviour, but not *vice versa*; being capable of autonomous action presupposes that one is capable of deliberative action, but not *vice versa*; and so forth. Most of those kinds of agency are distinctive of human agency, including acting for reasons. But none of them can be identified with human agency as such.

Is there a Problem of Disappearing Agents?

Michael Bratman and David Velleman are two contemporary proponents of the standard-causal model of agency, who have addressed a problem for the theory that is, apparently, related to the challenge of disappearing agency. In fact, some of their statements suggest that they acknowledge that there is a *problem* of disappearing agency. After introducing the central claims of the standard-causal theory in his paper 'On What Happens When Someone Acts', David Velleman identifies an apparently serious flaw of the model. It fails, as Velleman says,

> […] to include an agent—or, more precisely, [it] fails to cast the agent in his proper role. [According to the standard-causal model,] reasons cause an intention, and an intention causes bodily movements, but nobody—that is, no

> person—*does* anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part.[24]

Velleman, apparently, accepts the full force of the challenge of disappearing agency, and he acknowledges that it constitutes a problem, which the standard-causal theory must address and solve. In a fairly recent article, Michael Bratman addresses two interrelated and open problems for the standard-causal theory. One problem is what he calls the metaphysical problem of 'agential authority'. Introducing the problem, Bratman says that

> […] when a person acts because of what she desires, or intends, or the like, we sometimes do not want to say simply that the pro-attitude leads to the action. In some cases we suppose, further, that the *agent* is the source of, determines, directs, governs the action and is not merely the locus of a series of happenings, of causal pushes and pulls.[25]

In order to solve the problem of agential authority, Bratman suggests, the standard-causal theory must show that the phenomenon of 'agent- or self-determination' can be identified with or reduced to a complex 'causal structure involving events, states, and processes of a sort we might appeal to within a broadly naturalistic psychology.'[26]

Both Velleman and Bratman see themselves committed to the naturalistic approach. In other passages both make clear that they do not accept the full force of the challenge of disappearing agency. In fact, that is implicit in the quoted passage from Bratman, where he says that the problem arises only *sometimes*. Sometimes, he says, we want to say that the agent is the *source* of her actions, rather than just a locus of events and causal pushes and pulls. In other passages he acknowledges that the standard-causal model can capture what he calls 'merely motivated behaviour', which is, after all, a kind of agency. The problem of agential authority, then, does not arise with respect to all kinds of agency, but only with respect to what Bratman calls 'full-blown agency', in which the *agent* determines the action.[27]

That is very similar to Velleman's view. After saying that 'nobody *does* anything' on the standard-causal account, Velleman introduces the notion of 'human action *par excellence*'.[28] Velleman says he is interested in what distinguishes our

[24] Velleman, 2000, p.123.
[25] Bratman, 2001, p. 311.
[26] Ibid., p. 312.
[27] Ibid.
[28] Velleman, 2000, p. 124.

conception of *human* action from our conception of other animal behaviour, and that what he is trying to show is that the standard-causal theory

> […] describes an *action* from which the distinctively human feature is missing, and that it therefore tells us, not what happens when someone acts, but what happens when someone acts halfheartedly, or unwittingly, or in some equally defective way.[29]

Taken literally, this passage is self-contradictory. Velleman says that the standard-causal theory describes an *action*, yet it fails to provide an account of what happens when someone *acts*. I take it that an agent who acts defectively still performs an action. What Velleman must mean is that the standard-causal account is successful only in describing *some* actions—namely, somehow defective or sub-excellent ones—, but it fails to capture human action *par excellence*.

Velleman and Bratman think that they have identified a genuine metaphysical problem for the standard-causal model. The rhetoric used in some passages *suggests* that Velleman and Bratman acknowledge and address the problem of disappearing agency.[30] However, closer inspection shows that their primary aim is to reconcile the naturalistic standard-causal approach with a *special* and distinctively *human* kind of agency. The alleged problem is that the role that *agents* play in full-blown agency or action *par excellence* is not captured by the standard-causal model—we might call that the problem of disappearing agents.

Is there a problem of disappearing agents? What *exactly* is the problem? I will argue that there is no such problem. The alleged problem has something to do with the agent's role in the performance of 'full-blown' agency. In order to make progress, we first need to know what 'full-blown' agency—or 'human action *par excellence*'—consists in.

Identification and Autonomous Agency

Both Bratman and Velleman are interested in a higher and refined aspect of human agency, which they characterise, partly, by opposing it to lower, defective, or otherwise sub-excellent aspects of human agency. Both are influenced by the philosophy of Harry Frankfurt and Frankfurt-style examples, which feature, typically,

[29] Ibid., my emphasis.

[30] Berent Enç is a proponent of the standard-causal model who explicitly acknowledges that there is a problem of disappearing agency; see Enç, 2003, especially pp. 133-137. I will discuss some aspects of Enç's proposal in the next section.

the performance of an action that the agent does not want to do despite being motivated to do it. The best-known example of that kind is the 'unwilling addict' who is motivated to take a drug, even though he does not want to take it.[31] In Frankfurt's terminology, such an agent has a first-order desire to take the drug and a second-order desire that this first-order desire be not effective—by leading to action.

Generally, first-order desires are directed towards actions, objects or states of affairs, and second-order desires are directed towards first-order desires and their motivational efficacy. An agent *S* has a second-order desire either if *S* has a desire to have—or cease to have—a certain first-order desire (to do or attain something), or if *S* has a desire that a certain first-order desire be motivationally stronger or weaker than it actually is. In case the agent has a second-order desire not to be motivated by a certain first-order desire, as in the case of the unwilling addict, Frankfurt says that the agent cannot *identify* himself with the first-order desire; in other passages Frankfurt says the agent *dissociates* himself from having the desire or being motivated by it, and that the agent is *alienated* from it.[32]

Bratman and Velleman agree with Frankfurt that examples of that kind highlight an aspect of human agency that a comprehensive theory of action has to account for, and they agree that the notion of identification is appropriate to capture that aspect. But they also think that Frankfurt's own reconstruction of the phenomenon in terms of first- and second-order desires is seriously flawed and, ultimately, untenable. If that is correct, then the *problem* is to provide an account of the notion of identification in standard-causal terms—one that avoids the problems of Frankfurt's construal in terms of higher-order desires.

The problem with Frankfurt's proposal is the following. The model introduces second-order desires in order explain what it is for agent to identify himself with—or dissociate himself from—the desires that motivate his actions. It says, roughly, that to identify oneself with a first-order desire *is* to have the second-order desire that one be motivated by that first-order desire.[33] The problem is that it is possible to dissociate oneself—to fail to identify oneself—with the second-order desire in question, for

[31] See Frankfurt, 1971, reprinted in Frankfurt, 1988, as essay 2.
[32] See Frankfurt, 1988, essays 2, 5 and 12.
[33] In fact, that is only one possible interpretation of Frankfurt's early position, which he denied explicitly later. But since that is the reading that both Velleman and Bratman take as the starting point for their expositions, I shall assume it as well.

second-order desires are themselves *merely* desires. To account for the fact that one can identify oneself with the second-order desire, would require that there is a third-order desire to have the second-order desire, and so forth.[34]

Given that, *full-blown* agency or action *par excellence* is action that issues from desires with which the agent can identify herself, in the sense outlined. And the problem is to account for the notion of identification in a way that is compatible with the reductive standard-causal model. What is important to note, though, is that Frankfurt's primary aim was not to provide an account of identification. Rather, what Frankfurt was interested in was to provide an account of *autonomous* and *free* action and of the concept of a *person*. The account of identification was only instrumental with respect to those to objectives. According to Frankfurt's original position, whether an agent acts freely and autonomously depends on whether the agent can identify himself with the desires that motivate the action. And whether an agent is a person or not depends on whether the agent has or is disposed to have second-order attitudes towards her first-order desires—whether the agent, as Frankfurt says, *cares* about by which desires her actions are motivated.

Given that, it is a mistake to think the problem to account for *full-blown* agency or human agency *par excellence* constitutes a new and genuine metaphysical problem for the standard-causal theory. The problem to provide an account of identification is instrumental to solving traditional and well-known philosophical problems; namely, to specify the conditions for autonomous agency and personhood. Velleman and Bratman, we can conclude, neither address the fundamental worry expressed by Melden and Nagel, nor have they identified a novel problem in the metaphysics of agency.

Self-determination

The terms 'full-blown agency' and 'action *par excellence*' are not commonly used. It is not entirely clear to me whether Bratman and Velleman think that full-blown agency—or action *par excellence*—differs from autonomy in significant ways. The fact that they introduce those terms suggests so. What seems clear, though, is that full-blown agency has something to do with autonomy or self-determination.

[34] That was first pointed out by Watson, 1972. Compare also Bratman, 2001, p. 313 and Velleman, 2000, 132-135.

Velleman claims that the problem, which he is trying to solve, is the fundamental problem 'of finding a place for agents in the explanatory order of the world.'[35] When a human agent performs a human action *par excellence*, then the behaviour, as Velleman says, can be 'traced back to the agent himself rather than occurrences within him'.[36] Similarly, Bratman says that sometimes we want to say that the *agent* determines the action, rather than agent-involving states or events. Bratman aims to specify the structure of psychological mental states and events—or the 'psychological functioning', as he says—that is equivalent to a person's *being* the 'full-blown agent' of an action.[37] I think it does not matter whether *full-blown* agency—or action *par excellence*—differs from autonomy. Because, no matter what kind of agency is under consideration, it is misleading and unhelpful to characterise it in the way suggested by Velleman and Bratman; namely by referring to the *agent* is its source. Let me explain.

I argued that there is no such thing as agency *as such*. Agency is best understood as a family of certain kinds of agency that can be ordered as a hierarchy. Something similar, I think, holds with respect to agents. Velleman and Bratman think that some actions can be 'traced back' to the agent, and they seek to specify the agent's role in the production of such actions. But just as it is unhelpful to look for the phenomenon of agency *as such*, it is unhelpful to look for the *agent*—*him-* or *herself*—or the *agent's* role or functioning in the performance of action. Consider again the offered hierarchy of kinds of agency: purposive behaviour, acting for reasons, deliberative action, reflective and self-controlled action, free and autonomous action. Clearly, *in a sense*, each instance of behaviour, which falls under any one of those categories, *has* an agent—a being that performs the behaviour. *In that sense*, each instance of behaviour can be 'traced back' to an agent, and in each case we can specify the agent's role or participation. But that means that we cannot discriminate between different kinds of agency by asking whether the behaviour in question can be traced back to an *agent*. Maybe we can delineate human agency by asking whether a given action can be traced back to a *human* agent? But that cannot be right either, because most of the listed forms of agency are specifically *human* forms of agency—and they can, therefore, be traced back to a human agent. What we usually do in order to

[35] Velleman, 2000, p. 127.
[36] Ibid., p. 130.
[37] Compare Bratman, 2001, p. 312.

distinguish human forms of agency from other animal behaviour is to point out, or refer to, some of the *properties* or *abilities* of human agents by virtue of which they are capable of higher forms of agency—we do not refer to the fact that human actions can be traced back to an agent. To ask whether the action can be traced back to an *agent* or whether the agent plays a role in its performance strikes me as vacuous and inadequate. (Obviously, it would be uninformative and circular to say that full-blown actions or actions *par excellence* are actions that can be traced back to full-blown or excellent agents.)

One may object that Velleman and Bratman do not literally refer to the agent construed as a *substance*. Both Velleman and Bratman reject reference agent- or substance-causation, and they see themselves committed to the reductive approach of the standard-causal theory. Rather, when they talk about the *agent*, they mean some structure of psychological states and events that occupies the role of the agent.[38]

That observation is correct. However, it does not affect my criticism. Both Velleman's and Bratman's view can be understood as a continuation of Frankfurt's endeavour to account for free and autonomous action in terms—or *partly* in terms—of identification. The aim of that project is to identify a structure of psychological states and events, which constitutes an agent's identifying himself with his operative motives. If that reading of Velleman and Bratman is correct, then the problem they are dealing with is to provide conditions for free and autonomous action. To introduce the terms 'full-blown agency' and 'agency *par excellence*' is confusing and unnecessary.

Far more important, though, is the following point. Take any structure, *M*, of mental attitudes and relations between them, such that an agent's being in *M* constitutes or realises the agent's identifying himself with an operative motive (or, more generally, the exercise of a particular kind of agency).[39] To identify parts of *M*, or *M* as a whole, with the *agent* or the *agent's role* in the production of action, it seems to me, is a category mistake. If *M* constitutes or realises the agent's identifying himself with an operative motive, then it is the agent's *having*—or *being in*—*M* that constitutes or realises the phenomenon of identification (or the exercise of the kind of

[38] Compare Bratman, 2001, p. 312 and Velleman 2000, p. 137.
[39] Consider, for illustration, Frankfurt's original view, according to which an agent *S* identifies himself with a motive *m* if and only if *S* has a higher-order attitude that favours *m*'s presence and motivational efficacy.

agency in question). Given that, it is a mistake to construe the question whether the agent can identify himself with a motive as the question whether the *agent* stands in the *relation* of identification with some of his motives. We can say of the agent that he identifies himself with the motive, but we should not look for an agent *within* the agent who *does* that—we should not look for an entity that stands in the relation of identification with some of his attitudes.[40] In other words, it would be a mistake to ask whether identification can be 'traced back' to the agent, since identification is constituted or realised by the agent's being in the relevant structure of psychological states and events. Similarly, it would be a mistake to characterise full-blown agency—or action *par excellence*—by saying that it can be traced back to an agent, since autonomous agency issues from the agent's *being in* a psychological structure of a certain kind—and neither a part nor the whole of that structure is identical with the agent.

Deviant Causal Chains and Reason-Responsiveness

Most proponents of the standard-causal approach acknowledge that the possibility of deviant causal chains constitutes a serious problem and that the plausibility and success of the view depends on whether that problem can be resolved convincingly.[41] In this section I will first say more about the connection between the problem of causal deviance, acting for reasons, and agential control. Then I will introduce and discuss different forms of causal deviance, and, finally, I will propose a solution, which is based on guidance by and responsiveness to the contents of reason-states.

In the previous section we saw that the standard-causal theory can provide a viable account of an agent's exercise of control in terms of the agent's being guided by and responsive to reasons. It emerged that the problem of causal deviance is related to the issue of providing an account of acting for reasons and agential control. Examples such as Davidson's climber show that the fact that an agent's mental states and events cause *and* rationalise a bodily movement does not guarantee that the agent acts *for* reasons. Hence, the causal and rationalising relation is not sufficient for

[40] This mistake is particularly salient in Velleman's position. Velleman says that the problem with Frankfurt's hierarchical model of agency forces us to 'look for mental events and states that are functionally identical to the agent' (Velleman, 2000, p. 137).
[41] Compare for instance Brand, 1984; Bishop, 1989; Davidson, 1980; Enç, 2003; Mele, 2003; Searle, 1983; Thalberg, 1984.

acting for reasons. The same holds for action and agential control, because the standard-causal model construes action and control in terms of acting for reasons. Examples of basic deviance show that the fact that a movement is caused and rationalised by reason-states is sufficient neither for action, nor for an agent's exercising control. What is missing, as I suggested, is that the movement is *guided* by and *responsive* to the agent's reason-states. According to that suggestion, to amend the account of acting for reasons and agential control in the way outlined *is* to provide a solution to the problem of basic deviance. That means that causation in the right—that is, non-deviant—way is causation that meets the conditions of guidance and reason-responsiveness.

Causal Deviance

It is common to distinguish between three different types of deviant causal chains in the theory of action. Davidson's climber example, which has been introduced above, is a case of *basic* deviance, which is the most troublesome kind of causal deviance. Besides that we can distinguish between *consequential* deviance and *second-agent* deviance.[42] The former is a second main category of causal deviance—along with basic deviance—, whereas second-agent deviance can be treated as a special case of basic deviance. In order to distinguish these two further kinds, I need to say more about basic deviance.[43]

In all cases of deviance, some control-undermining state or event occurs between the agent's reason-states and an event produced by that agent. What distinguishes cases of basic deviance is that the control-undermining event occurs *between* the reason-states and an agent-involving event that would constitute a *basic* action.[44] Davidson's climber example is a case of basic deviance. The climber rids himself of the weight and danger by loosening his hold on the rope. If the climber performed an action, the movement of loosening his hold on the rope would be a basic act. The state

[42] I borrow the term 'basic deviance' from Bishop, 1989, p. 132, and the term 'consequential deviance' is taken from Brand, 1984, p. 23.
[43] I may not do justice to all the subtleties of the phenomenon of deviance in action theory. It may be possible to introduce some further sub-categories that require modified responses. The exposition, though, covers the standard cases of deviance, and with basic deviance it covers the most important and most troublesome kind of causal deviance. I take this to be sufficient to show that the problem of causal deviance can be resolved convincingly.
[44] For the notion of *basic* action see chapter 1, pp. 18.

of nervousness that undermines the agent's control occurs between the climber's reason-states and that movement, which would be a basic action.

In the case of *consequential* deviance, the control-undermining states or events occur somewhere between a basic action and some action—or outcome—that the agent intended or wanted to perform—or bring about—*by* performing the basic act. Consider the following standard example. A sniper has the intention to kill an enemy by shooting him. He carries out the intention, but misses. By producing the noise of the shot, though, he stampedes a herd of wild pigs, which trample the poor enemy to death.[45] In such cases the agent performs a basic action intentionally, for reasons, and the agent is in control as far as that basic action is concerned. The deviance occurs later and it affects an action or outcome that the agent wanted to bring about *by* performing the basic action. These cases are thought to constitute counterexamples, because the agent's end is *not* to perform the *basic* action. The sniper's end is to kill the enemy; he intends to do so, and the intention causes the enemy's death. According to a *simple* standard-causal theory, the sniper does not only perform the action of firing the shot intentionally and for reasons, but he also kills the enemy intentionally and for reasons. The former is true, but the latter, it seems, is clearly false. Hence, the standard-causal theory has got it wrong.

Cases of *second-agent* deviance are special cases of basic deviance, as control is undermined by interference between the agent's reason-states and the movement that would be a basic act. Second-agent deviance occurs when control is undermined not by states or events, but by another *agent*. Harry Frankfurt has presented a much-discussed example, which fits exactly this description, in order to make a point about moral responsibility.[46] The agent, Jones, is about to decide whether to vote for one of the two presidential candidates. Unbeknownst to Jones a gifted neuroscientist, called Black, has implanted a device in Jones' brain by virtue of which he can detect what Jones is going to decide and influence his actions, if he wishes to do so. Jones, it seems, it not *fully* in control, no matter whether Black intervenes or not. And it seems that the causal pathway is deviant—at least in those cases in which Black intervenes.

[45] Compare Bishop, 1989, p. 126, who has taken this example from Daniel Bennett. For another much discussed example of that kind see Chisholm, 1966.
[46] Frankfurt, 1969, reprinted in Frankfurt, 1988, essay 1.

Let us consider first a standard way of dealing with consequential deviance. What goes wrong in cases of consequential deviance is that the intended action or outcome is not brought about according to plan. The sniper intends to kill the enemy in a certain way; namely, by shooting him, not by having him trampled to death. The standard-causal model can accommodate cases like that by requiring that actions are guided by the contents of the relevant reason-states. In cases of consequential deviance, the agent has a certain action-plan that is incorporated in the contents of the relevant reason-states. In the example, the action-plan consists in the intention to kill the enemy by shooting him. To say that the performance of the action must be guided by the intention is to say, simply, that the way in which the intended end is brought about must be in accord with the agent's action-plan, which is incorporated in the content of the intention.[47] The notion of *guidance* can be accounted for in causal terms, if it is granted that the intentional contents of reason-states and events can be causally relevant. Guidance by reason-states can then be construed as causation by reason-states *in virtue of* their contents. A bodily movement, for instance, is in that way guided by reason-states, because the reason-state's content is causally relevant as to whether *that* particular movement occurs, rather than another one or no movement at all.[48] Requiring guidance by reason-states, the standard-causal model can exclude cases of consequential deviance.[49]

---

[47] For a summary account of the role of action-plans in intentional action see, for instance, Mele and Moser, 1997.

[48] Consider, for instance, guidance by an intention—by an agent's having or acquiring an intention. The simplest form of an action-plan, which is part of the content of an intention, consists of a representation of the intended action; a representation of which *type* of action is to be performed. In most cases, though, the action plan specifies by what means—by the performance of which type of act—the intended or desired end should be attained. An act-token is then guided by an intention if, firstly, it either instantiates the intended act-type or if it instantiates an act-type that is specified as a means to the intended end, and, secondly, if it is caused by the intention in virtue of its content—which incorporates the action-plan.

[49] Many have argued that the causal connection between reason-states and the action must be *sustained* or *continuous*—at least in some cases where the agent's exercise of control is a process that takes as long as the performance of the action itself. Compare for instance Bishop, 1989; Mele, 2003; Thalberg, 1984; Searle 1983. The standard way of accommodating such cases is to say that the *guidance* function of intentions—or reason-states in general—can be sustained or continuous. That requires that the agent *is having* the intention as long as the execution of the action takes—which seems unproblematic. The causal relation between such an intention and the action might be construed as *sustained* or *continuous* causation. But it can, more plausibly, also be construed as a series of *feedback loops* between behaviour and intention. That means, very roughly, that the causal pathway would incorporate a mechanism that adjusts, at certain time-intervals, the execution of the action in accordance with the content of the intention. Compare for instance Bishop, 1989, pp. 170-171 and Mele 2003, pp. 56-58. Further, cases in which the agent must *improvise* cause complications—either because the plan is not

To refer to the guiding role of the contents of reason-states helps only when the agent has an action plan that covers the event that would have been an action, had the pathway not been deviant. That is why it helps only in cases of consequential deviance. Consider again Davidson's climber. He wants to rid himself of the weight and danger and believes that loosening his hold on the rope is a way of doing so. He does not have a further belief how to loosen his hold; he does not have a belief concerning what means are necessary, appropriate or best for loosening his hold, because loosening the hold is a basic act. We do not need to *plan* basic actions, because we do not have to know how or by what means to bring them about. We just do them; that is why they are basic. And that is why reference to guidance by reason-states does not help with basic deviance.[50]

It has been pointed out that an action that is done *for* a reason is a *response* to that reason; action that is motivated by reasons is *responsive* or *sensitive* to reasons.[51] In the previous section I have outlined how the requirement of reason-responsiveness can help to handle examples of basic deviance. I will argue now, firstly, that the problem of basic deviance can in fact be solved by requiring reason-responsiveness, and, secondly, that the notion of reason-responsiveness is compatible with the standard-causal model.

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 relevant events in that example are following: the agent's having the reason-state (say, the intention to rid himself of the weight and danger), the agent's being nervous, and the agent's movement of loosening the hold on the rope; call these events *r*, *n* and *m*. In the actual scenario *r* causes *n*, and *n* causes *m*. What will be relevant to an explanation of why the causal pathway is deviant are the following three facts. Firstly, *m* is caused by an event, *n*, which undermines the agent's control. Secondly, *n* does not rationalise *m*, but is caused by an event that rationalises *m*. And thirdly, the fact that *m* is caused *and* rationalised by an appropriate reason-state, *r*, is a coincidence—because it is a

specific enough or because things are not going according to plan. Brand, 1984, and Bishop, 1989, show that the guidance condition can be refined as to accommodate such cases.

[50] Compare Bishop, 1989, pp. 132-134.

[51] Compare, for instance, Audi, 1997; Bishop, 1989; Peacocke, 1979; Stoutland, 1998b.

coincidence that *n* causes the type of movement that is rationalised by *m*. Basic deviance is possible because coincidences of that kind are possible.

Proximate Causation

A first strategy to solve the problem of basic deviance is to require proximate causation. It says that intentions are the proximate or immediate causal antecedents of all actions, and it excludes control-undermining states by excluding all causal intermediaries between reason-states and actions.[52] (Note that it would be problematic to exclude only control-undermining states and events, because we must explain agential control in terms of non-deviant causation.)

Bishop has dismissed that strategy for the following reason. Intentions, Bishop claims, 'as realised in central neural states will never be causally proximate to the bodily movement that matches them'.[53] There will be a causal chain of physiological events that results in the movement, and the most the proximity strategy can require is that the intention initiates that chain. But that, of course, does not solve the problem, because that causal chain may run through a state or event that undermines control.

That objection, however, is flawed as it confuses levels of explanation. The requirement that intentions must be proximate causes of action is a requirement at the level of psychological description and explanation. It may well be—and it is almost certainly the case—that a causal relation between two intentional events is realised at the neuro-physiological or physical level by a far more complex chain or pattern that involves a multitude of events. But that does not show that the intention cannot be the proximate *mental* antecedent of action.[54]

The proximity strategy, however, is unsatisfying in one important respect. I said that the problem in the climber case is that the bodily movement is merely caused by the reason-state, rather than being a response to it. The proximity solution ensures that bodily movements, for instance, are caused and rationalised in a way that constitutes agential control. But it does not show that the bodily movement is a *response* to the reason-state *qua* reason-state; it does not show that the reason-state causes the bodily movement *because* it is a *reason*-state. The reason-state causes *and* rationalises the action, but the reason-state's rationalising the action seems to be irrelevant to its

[52] Compare, for instance, Brand 1984, Mele and Moser, 1997.
[53] Bishop, 1989, p. 139.
[54] Compare Wedgwood, *forthcoming*, section 3.

causing it. The proximity strategy establishes agential control by excluding all causal intermediaries—which excludes, trivially, all control-undermining intermediaries. But it does not establish control as reason-responsiveness.[55]

The Counterfactual Strategy

A second strategy for solving the problem of basic deviance attempts to capture the notion of reason-responsiveness in counterfactual terms. Let us consider again the climber example and assume that in the actual scenario the climber has the intention of ridding himself of the weight and danger by loosening his hold on the rope *now*—at time *t*. And consider a possible scenario in which the climber intends to rid himself of the weight and danger by loosening his hold on the rope not now, but *shortly*—at *t*'. Other things being equal, that intention would unnerve the climber just as in the actual scenario, and we can assume that the climber's nervousness would cause the loosening of the hold on the rope *now*—just as in the actual scenario. We can see, then, that the agent's movement in the *actual* scenario is *not* reason-responsive, in the sense that the climber would *not* have performed a different action, had he had a slightly different reason-state—an intention that calls for a performance of the loosening of the rope at *t*' rather than at *t*. Given that, it seems that an appropriate counterfactual condition can solve the problem of basic deviance. Consider as a first and rough approximation the following condition, taken from Berent Enç's recent treatment of the problem of causal deviance.

> (CC) If the content of the intention had been different, the action would have been different correspondingly.[56]

The required *correspondence* between the intention's content and the action consists in a match between the type of the basic act performed by the agent and the type of basic act specified in the content of the intention. So, had the agent intended to *B*, rather than *A*, then the intention would have caused a bodily movement that is, or constitutes, an action of type *B* (and had the agent intended to bring about the end *E* by performing the basic act *B*, rather than by performing *A*, the agent's intention

[55] Another worry is that the proximity strategy is *ad hoc*, as it simply excludes what is responsible for the problem. But there is independent plausibility to the claim that all actions are preceded and accompanied by the agent's having or forming an intention. Searle, for instance, argues on phenomenological grounds that every overt action consists of a bodily movement and what he calls an 'intention-in-action' that causes the movement (1983, chapter 3).

[56] Compare Enç, 2003, p. 103, who adapts the condition from Bishop, 1989, p. 150.

would have caused a bodily movement that is, or constitutes, an instance of *B*, rather than *A*). Had the climber intended to rid himself of the weight and danger *shortly*, the nervousness would nevertheless have caused his loosening the hold *now*. That is, his action would not have been different correspondingly. The counterfactual condition is not satisfied, and the climber case fails, therefore, to constitute a counterexample.[57]

Note that the counterfactual strategy captures the notion of reason-responsiveness satisfactorily. Reason-states are rationalising states; they are states in the light of which the performance of some action appears as intelligible. In particular, the performance appears as intelligible in the light of the *contents* of reason-states. If Sue intends to *A*, then her *B*-ing appears as intelligible in the light of the fact that Sue believes that *B*-ing is conducive to *A*-ing—Sue's *B*-ing appears as intelligible in the light of the content of that belief. CC requires that actions co-vary with the *contents* of intentions, and that captures the intuition that actions, which are done for reasons, are responsive to reason-states *qua* reason-states.[58]

The problem with that approach is that it is conceivable that an agent is reason-responsive only in the actual scenario. Consider an agent, *S*, whose arm is replaced by a prosthetic device that is suitably connected to nerve endings of *S*'s nervous system. The only movement that the device can perform—at the time *t*—is a movement of type *M*. Assume that *M*-ing is intrinsic to *A*-ing and that *S A*-s at *t* because *S* intends to attain *E* by *A*-ing at *t*. Had, for instance, *S* intended to attain *E* by *B*-ing, then the action would not have been different correspondingly, under the assumption that *B*-ing requires a different movement of the prosthetic device. That example does not satify CC, but *S*'s attaining *E* by *A*-ing at *t* may nevertheless be reason-responsive: whether or not the action in the actual scenario is reason-responsive is independent of

[57] Reason-responsiveness certainly does not require that CC holds for all possible worlds. The climber example suggests that the relevant worlds are *close* worlds in which the motivational component of the climber's intention and part of its content are being held constant (compare Bishop, 1989, pp. 148-150). Further we held constant the circumstances and the climber's disposition to get nervous in the circumstances. Compare Haji, 1998, who specifies which features must be held constant in those alternative scenarios that are relevant to agential control (see pp. 80-82).

[58] Bishop, 1989, and Enç, 2003, think that the counterfactual strategy is subject to counterexamples in which another agent undermines the agent's control in the relevant alternative scenarios. An example of that kind is the above-mentioned Frankfurt example (Frankfurt, 1988, essay 1). However, the counterfactual strategy can avoid such counterexamples by holding constant the causal mechanism that is operative in the actual scenario and by changing the condition as follows: had the content of the intention been different and had the *actual mechanism* been operative, the action would have been different correspondingly (compare with the discussion of Frankfurt-style examples in Fischer and Ravizza, 1998). The remaining difficulty, which is not insurmountable, is to specify in virtue of what a mechanism is the same or a different mechanism.

whether the operative causal mechanism responds to reason-states in alternative scenarios.

Causation in Virtue of Content

The third strategy is a response to the original objection. In the climber example the reasons *merely* cause *and* rationalise the performance of the bodily movement. What renders the pathway deviant is the state of nervousness, which undermines the agent's control and renders it a coincidence that the reason-states cause and rationalise the bodily movement. What has been overlooked, though, is that the standard-causal model not only requires that mental states rationalise and cause actions, but also that they causally *explain* them. In the example, the reason-states explain the movement in some sense—but not in the sense required. The reason-states cause the nervousness, and the nervousness causes the movement. In the circumstances, had the agent not had the reason-states, the bodily movement would not have been performed. In that sense, the reason-states are causally explanatory of the movement. However, the theory requires that reason-states cause and causally explain actions in virtue of their intentional content. That requirement is clearly violated in the climber example. The statement of the problem presupposes that the relation of causation is transitive. The reason-states cause the movement, because they cause the nervousness, which causes the movement. But the reason-states do not cause the movement in virtue of their content, because the nervousness, trivially, does not cause the movement in virtue of content. In that case, the question of whether the relation of causation *in virtue of content* is transitive does not even arise, because the nervousness does not cause the effect in virtue of content. Accordingly, the reason-states do not explain the occurrence of the particular movement in virtue of their content—why *that particular type* of movement occurred, rather than another, cannot be explained by reference to the contents of the reason-states. Subsequently, the question of whether the relation of being explanatory in virtue of content is transitive does not arise, because the nervousness is not explanatory in virtue of content. In other words, if we consider the pathway from the reason-states to the movement, the relation of causation holds and it is transitive, but both the relation of causation in virtue of content and the relation of being explanatory in virtue of content break down.

That solves the problem of basic deviance and it captures the notion of reason-responsiveness. Being caused and causally explained in virtue of content, the action is

not merely a response to a cause, but it is a response to a reason-state *qua* reason-state; it is a response to the content of the mental state in the light of which its performance appears as intelligible.

Conclusion

I outlined a solution to the problem of consequential deviance and I presented three different solutions to the problem of basic deviance—which are three solutions to second-agent deviance, as I assumed that second-agent deviance is a special case of basic deviance. I expressed reservations with respect to the first and the second proposal. It is important to note, though, that the three proposals are not exclusive; they are not incompatible and they are not rival proposals. In particular, a standard-causal theory may require that reason-states proximately cause and causally explain actions in virtue of their content. That solves the problem of basic deviance and it captures the notion of reason-responsiveness.

Acting for Reasons and Deliberative Action

As it stands, the standard-causal theory says that non-deviant causation by reason-states is necessary and sufficient for acting for reasons. Some philosophers think, though, that acting for reasons requires that the action is based on—or that it results from— practical deliberation. They think, in other words, that all acting for reasons is, necessarily, *deliberative* acting for reasons. In this section I will discuss and reject Berent Enç's arguments for that position.

Enç is a proponent of the standard-causal approach. According to his theory, every action is based on an intention, and every intention is based on a practical deliberation in which the agent considers alternative courses of action and assesses the reasons for and against them. The functioning and realisation of that process, as Enç shows, can be construed in standard-causal terms only.[59]

Now, it is certainly not obvious that acting for reasons presupposes deliberation. Quite to the contrary, one may argue on intuitive or phenomenological grounds that Enç's reconstruction of what acting for reasons consists in is obviously false. Experience tells us that on many occasions when we act for reasons, we do not engage in deliberation. We can certainly distinguish between actions that are merely

[59] Enç, 2003, especially chapter 5.

done for reasons and ones that are preceded by an evaluative process in which one assesses or weighs the reasons for and against certain courses of action. Often it is obvious what one should do, often we do it out of habit, and sometimes there is simply not the time to assess the pros and cons. In all such cases our actions are not based on deliberation, but we may nevertheless act for reasons. Such considerations are not decisive. But they show that the proponents of the deliberative model must motivate their view, as it lacks intuitive appeal.

Purposive Behaviour and Acting for Reasons

What motivates Enç to defend a deliberative model of acting for reasons? Interestingly, Enç thinks that the theory provides a response to what I have called the challenge of disappearing agency. He says that a hard 'problem for a causal theorist of action is to persuade the sceptic that a coherent concept of *agency* or *control* can be located in mere event causation.'[60] According to Enç, that problem can be formulated, or expressed, in two ways. The first formulation is basically the one that we have encountered discussing Melden and Nagel:

> If an agent's action is the causal consequence of a series of events, then what sense can one give to the notion of that the action's being under the agent's control?[61]

The second formulation uses a comparison between acting for reasons and a less sophisticated form of behaviour. Enç considers the dive performed by a moth in order to escape a predator. The behaviour, we assume, is *caused* by the reception of an input, which indicates the advance of a predator, in conjunction with the moth's disposition to perform a dive in the circumstances. Comparing the two kinds of behaviour, Enç says the following:

> Certainly, the moth does not have the proper kind of *choice* or *control* over its dives. So the causal theorist ought to be able to say what it is that separates rational agents from the moth. And apart from pointing to the complexity of the mechanism in rational agents, which is just a quantitative difference, not a qualitative one, the causal theorist cannot in principle have any resources with which to do this.[62]

[60] Ibid., p. 133.
[61] Ibid.
[62] Ibid., p. 136.

The behaviour of the moth, as Enç points out, is purposive. The moth performs the dive *in order to* escape predation. But the moth, of course, does not dive *for* that reason. Enç thinks that this difference is a *qualitative* one, and he aims to account for it in standard-causal terms only.

Without going into much detail, I will in the following outline some of the central features of Enç's causal theory of deliberation. Then I will argue that Enç's aim and motive to account for control and for the qualitative difference between purposive behaviour and acting for reasons does not support the claim that all acting for reasons is based on deliberation. (In the following one must bear in mind that according to Enç an account of the kind of control exercised by rational agents and an account of the qualitative difference between purposive behaviour and acting for reasons is one and the same thing.)

Enç's Causal Theory of Deliberative Action

According to Enç, the crucial difference between acting for reasons and purposive behaviour is that an agent who acts for reasons has beliefs of a particular kind, which play a distinctive causal role in the process that leads to action. The beliefs in question have a conditional content of the form 'if under certain circumstances a certain type of action is performed, it is likely that a certain type of result will obtain'. What plays the mentioned distinctive causal role are the *consequents* of such conditional beliefs; for an agent to act for reasons, the consequents must, as Enç says, enter in into the causal mechanism as separate units.[63]

This sub-theory concerning the detachability and efficacy of the consequents of conditional beliefs is at the heart of Enç's causal theory of deliberative action. What motivates that sub-theory? Enç's aim is to account for control and for the qualitative difference between purposive and rational behaviour. If that aim *requires* the outlined theory of deliberative action (including the sub-theory), then there is good reason to endorse it. Furthermore, if an account of control and acting for reasons requires the outlined theory of *deliberative* action, then we have reason to endorse the view that all acting for reasons is based on deliberation.

Let us first see why Enç thinks that a causal theory of deliberative action requires the outlined sub-theory. Enç presents a sophisticated causal model of deliberative

[63] Compare ibid., p. 145.

acting for reasons that accounts for both deliberation about alternative *means* to an end and for deliberation about alternative *ends* (or courses of action). A central element of that model is what Enç calls the 'What-If generator'.[64] An agent who deliberates asks himself practical questions such as "What should I do in such and such circumstances?" or "How should I bring such and such about?" Asking practical questions, the agent will, according to Enç, run a series of what-if scenarios. That is, the agent will ask himself "What will happen if I *A* in the circumstances—what if I *B* instead?" and so forth. In order to come to a practical conclusion, the agent has to assess the value of the expected consequences and the costs of the pursuit of the required means. *A*-ing, for instance, may have better consequences than *B*-ing, but *A*-ing may be very costly or risky.

An evaluative process of that kind requires information—a background that allows the agent to compare and assess the alternatives. This information is provided by beliefs concerning actions and their consequences, and the content of such beliefs will say that 'if under certain circumstances a certain type of action is performed, it is likely that a certain type of result will obtain.' Deliberation, then, requires the kind of conditional beliefs introduced by the sub-theory. Further, the process of going through the What-If scenarios requires that the agent is able to *compare* the expected consequences of alternative courses of action. And in order to do that, the agent must be able to *detach* the consequent of the conditional belief. Given all that, we can see that a reconstruction of the whole deliberative process in standard-causal terms may require that detached consequents (of conditional beliefs) enter in into the causal mechanism as separate units.

So, if Enç's reconstruction of deliberative action for reasons is correct, then there is reason to endorse the sub-theory. But on that ground there is, obviously, no reason to think that acting for reasons and exercise of control is necessarily preceded by deliberation about alternative means and courses of action.

Enç, it seems, thinks that the aim of accounting for control gives us reason endorse the claim that all acting for reasons is based on deliberation, because the detachability and efficacy of a conditional belief's consequence is a feature that establishes a *qualitative* difference between acting for reason and merely purposive behaviour. But that alone, if true, does not support the claim that all acting for reasons

[64] See ibid., p. 157-159.

is deliberative. What one would have to show, in addition, is that *only* the proposed deliberative model can account for the qualitative difference in standard-causal terms. But I think it is easy to show that there are other differences between acting for reasons and purposive behaviour that count as qualitative differences.

Note, first of all, that Enç does not properly justify the claim that an account of the difference between purposive behaviour and acting for reasons must identify a qualitative rather than just a quantitative difference. Further, he does not explain what a qualitative difference—as opposed to merely quantitative difference—actually amounts to. But let us assume, for the sake of the argument, that we must identify a qualitative difference, and let us rely on an intuitive grasp of what a *qualitative* difference is.

Very plausibly, Enç's theory does identify a qualitative difference. Nothing in the causal history of the moth's behaviour plays a causal role similar to the one of detached consequents (of conditional beliefs). It identifies a *kind* of causal role or mechanism, which is operative in rational agents, but not in agents who are capable merely of purposive behaviour. But is that the only qualitative difference we can identify?

To begin with, a moth does not have propositional attitudes such as beliefs, desires or intentions. That *is*, it seem obvious to me, a qualitative difference between the two kinds of agents. And it is a difference that has significant implications with respect to what kinds of agency they are capable of. The fact that the moth does not have propositional attitudes entails, trivially, that its behaviour is not caused by *reason*-states, and that it cannot appear as reasonable or intelligible in the light of such states. Further, it entails trivially that the moth's behaviour is not *guided* by reason-states (that it is not caused in virtue of their contents). And the fact that the moth does not have beliefs entails, trivially, that nothing in the causal history of the moth's behaviour plays a causal role similar to the one of conditional beliefs.

Now, I cannot see why only that last difference should count as a qualitative difference. As far as I can see, all the differences stated in the previous paragraph are *qualitative* differences between merely purposive behaviour and action that is done for reasons. Given that, the aim of accounting for that qualitative difference does not lend any support to the claim that all actions that are done for reasons are based on deliberation. Since we do not *need* to appeal to deliberation and the causal role of

conditional beliefs in order to identify a qualitative difference, there is no need to assume that acting for reasons is necessarily deliberative.

Is Acting for Reasons Necessarily Based On Deliberation?

In the passage quoted above Enç says that 'the moth does not have the proper kind of *choice* or *control* over its' behaviour. Taken literally, Enç assumes that having control over something is the same thing as having a choice about it. If we assume further that making a choice is the same as deliberately deciding and acting for reasons, then Enç would simply be begging the question. Because then it is assumed that an account of control *is* an account of *deliberative* action done for reasons.

There is, however, an intuition that gives independent support to the assumption that acting for reasons requires some kind of deliberation. When we act for reasons, the reasons do not lead to action in an unmediated or straightforward way. Strictly speaking, it is not the reason-states (or the agent's having the reason-states) that result in actions, but the agent acts on them—in the light of them. Whenever an agent acts for reasons, the agent considers the reasons and, then, acts on them. Acting for reasons involves this minimal reflective element: the agent stands back and considers the action and its consequences. This minimal reflective process is often performed or executed in an instant, without hesitation, critical evaluation or reflection. Acting for reasons, the agent must consider at least two alternative courses of action and their consequences. But sometimes that will involve simply considering doing one thing and the consequences thereof, and considering the consequences of not doing it.[65]

That intuition has some plausibility. Probably there is a good case to be made for the claim that we engage in deliberation far more often than one might think—especially if one construes deliberation as a process that is often carried out swiftly without hesitation and without much reflection or evaluation. However, I do not think that the outlined intuition gives us decisive reason to think that all acting for reason is necessarily based on deliberation. As pointed out, there is the intuition to the contrary that there are reasonable or rational actions—actions done for reasons—that are not preceded by any deliberation or reasoning concerning alternative possibilities—not even by a minimal or very brief process of deliberation. Alfred

[65] Compare, for instance, Schueler, 2003, who argues on intuitive and conceptual grounds that acting for reasons is necessarily deliberative acting for reasons. Compare also Enç, 2003, pp. 153-157.

Mele, for instance, argues that some intentions are acquired *passively* in the sense that they are not actively formed as the result of a deliberative process. Habitual actions provide good examples. 'When I intentionally unlocked my office door this morning', Mele says, 'I intended to unlock it. But since I am in the habit of unlocking my door in the morning, and conditions this morning were normal, nothing called for a *decision* to unlock it.'[66]

Mele argues further that some intentions 'nonactionally arise out of desires'. Consider Al who is on the way back to this office. Suddenly, Al acquires the desire to include a short detour—he just feels like taking a short detour. Mele suggests that it is plausible to say that such a desire straightforwardly results in the intention to take a short detour (which straightforwardly results in Al's taking a short detour) provided that Al has no relevant competing desires and no reservations concerning the desired action. Accordingly, the standard-causal theory might require that an agent acts for reasons by passively acquiring an intention and performing an action 'out of a desire' only if there are no competing desires and only if the agent has no reservations (in the form of beliefs that the desired course of action is imprudent, unethical, too risky, or the like).

Alternatively, proponents of the standard-causal theory may require that an action that is motivated by a desire is done for reasons only if the desire is *integrated* in the agent's motivational system, in the sense that it is responsive to opposing reasons. Robert Audi has formulated such a condition on acting for reasons.[67] When an agent acts for reasons, opposing reason will at least *diminish* the agent's tendency to perform the action in question. And it will be true that had the agent had opposing reasons, they would have reduced that tendency. For instance, if an agent *S* has a tendency to *A* because *S* desires *A*-ing, and if *S* has no competing desires, then the desire to *A* is *integrated* into *S*'s motivational system in case it is true that if *S* believes or judges that *A*-ing is in some sense inappropriate, wrong or undesirable, then that belief or judgement will diminish *S*'s tendency to *A*.

At this point a remark concerning the role of desires seems in place. One thought or intuition that stands behind many of the objections and worries we encountered is that being *caused* to do something by a desire is indistinguishable from being *driven*

[66] Mele, 2003, p. 200.
[67] Compare Audi, 1986, especially pp. 93-95.

or *pushed* by a desire. The first thing to say is that desires are not *per se* irresistible or overwhelming. We can distinguish between being caused to do something by a generic desire and being caused to do something by an irresistible or overwhelming desire. The important point is that what renders a desire irresistible or overwhelming is *not* the fact that it *causes* an action.

Enç and proponents of Frankfurt's hierarchical model of agency would certainly agree with that. But they offer an additional criterion for acting on ordinary—that is, resistible or non-compulsive—desires. On the hierarchical approach, the agent must identify himself with the desire by virtue of having higher-order pro attitudes towards it. Enç, on the other hand, requires, firstly, that the agent has at least one other current desire favouring an alternative course of action and, secondly, that the agent considers acting on either desire in a process of deliberation.[68]

I do not deny that those proposals characterise interesting and important aspects of human agency. What I question, however, is that deliberation (or identification) is *necessary* for acting for reason (and the associated kind of control). I showed that we can make the relevant distinctions within the standard-causal framework without referring to either deliberation or identification. As with respect to desires, we want to distinguish between acting on a desire and being driven or pushed by a desire. We do not need to refer to either deliberation or identification, since we can account for that difference by referring to relations between desires and other *first-order* attitudes of the agent, such as competing desires or beliefs about what is appropriate, good or desirable. On that suggestion, a desire is irresistible or overwhelming if it is not in *accord* with and not *responsive* to opposing reason-states. Action that issues from such unresponsive desires does not constitute action for reason, and a causal pathway that leads from unresponsive desires to action does not constitute an exercise of control, even if the desire rationalises and causes a movement in a non-deviant way.

What these suggestions show is that the standard-causal theory has the resources to make all the relevant distinctions. It can distinguish between purposive behaviour and acting for reasons, and it can distinguish between behaviour that results from irresistible or overwhelming desires and action that is based on desires. What emerges is a plausible account of acting for reasons, which does not require that acting for reasons is based on deliberation.

[68] Enç, 2003, p. 160.

Finally, let us consider Enç's account of habitual action. Enç is well aware of the fact that habitual action poses a challenge to any view that claims that all acting for reasons is deliberative. Enç considers an agent who 'always removes the key from the ignition before he opens the door of his car', and he acknowledges that in cases like that 'no weighing of pros and cons takes place before the habit kicks in'.[69] Enç does not claim, as one might expect, that in the case of habitual action we are not aware of, or do not attend to, the deliberative process (because the deliberation is processed in an instant or because it is sub-conscious). Rather, Enç suggests that a habitual action is done for reasons if the habit, which is enacted in the performance of the action, has been acquired through a process of deliberation.

> […] in so far as the circumstances are perceived to conform to those in which the habitual action was adopted through deliberation as the best means of reaching some goal, and the goal is still a relevant concern, enacting the habit is a piece of rational behaviour.[70]

So, in fact Enç does not require that every action that is done for reasons is the direct or immediate result of a deliberative process. Rather, he requires that every rational action is grounded in—and can be traced back to—a deliberative process in the sense just outlined.

But that approach to habitual action gives rise to a serious problem. Recall that Enç aims to meet the challenge of disappearing agency—the problem of 'agent-control', as he calls it—by providing an account of acting for reasons in standard-causal terms. Acting for reasons, he thinks, presupposes deliberation, which means, in effect, that exercise of control presupposes deliberation. In conjunction with the suggested view on habitual action we obtain the following problem.

According to the view, whenever an agent performs a rational action habitually at the time *t'*, then the agent performed a deliberative action for reasons at some earlier time *t*, which resulted in the formation of the habit in question. It is not implausible to think that the rationality of the habitual action is grounded in—and can be traced back to—the deliberative action at *t*. What is rather implausible, though, is that the exercise of control at *t'* can be traced back to the deliberative action at *t*. Control, it seems

[69] Enç, 2003, p. 154.
[70] Ibid.

obvious, is exercised at the time the action is performed. Since control is construed as deliberative action for reason, we must conclude for the case of habitual action that the agent exercises control only at *t* when then habit is acquired, but not at *t*' when the habit is enacted. And that is, I think, a counterintuitive and very unfavourable consequence of Enç's theory.

In conclusion we can say that Enç has failed to motivate the claim that deliberation is necessary for acting for reasons. Hence, he failed to give us reason to think that deliberation is necessary for an agent's exercise of control. Enç argues, in effect, that acting for reasons and deliberative action collapse into one category. However, if my arguments are correct, the distinction between acting for reasons and deliberative action does not collapse; the distinction is significant and real. It may well be that many—or even most—instances of actions that are done for reasons *are* based on deliberation. But the question is not whether some or most instances of rational action are based on deliberation, but whether rational action is necessarily deliberative. It is conceivable that we act for reasons without engaging in any deliberation, and intuition tells us that occasionally we do perform such non-deliberative rational actions. I argued that we do not need to abandon that intuition.[71]

Practical Reasoning and Treating as a Reason

Given that deliberation is not necessary for acting for reason, it does, of course, not follow that non-deviant causation by reason-states is sufficient. One may think, as already mentioned, that acting for reasons involves a minimal reflective element of treating a certain consideration as a reason for action. One may endorse that intuition and deny that treating as a reason requires deliberation about alternatives courses of action. There are two further interpretations of that intuition that would support the claim that non-deviant causation by reason-states is not sufficient.

According to the first, treating as reasons involves practical reasoning about the means to a given end. I distinguished between acting for reasons and acting for

[71] The task of settling the question whether acting for reasons is necessarily deliberative action strikes me as difficult and elusive. Our intuitions concerning the concept of acting for reasons are neither straightforward nor conclusive. Further, a proponent of the view that acting for reasons is necessarily deliberative can always avoid counterexamples by claiming that the agent need not explicitly or actively engage in a deliberative process or that this process takes place at a sub-conscious level. I think the suggested approach to ask whether we need to refer to deliberation in order to obtain all the relevant distinctions and in order to obtain a plausible characterisation of acting for reasons is the best way to approach the issue.

normative reasons. The former is action that can be rationalised in the light of the agent's reason-states, the latter is action that is done for good reason. I pointed out that non-deviant causation by reason-states is supposed to capture the former rather than the latter. However, the notion of acting for reasons is ambiguous in a further sense. Setting aside the question of whether the agent in fact had good reasons, we can ask whether the agent *took herself* to have good reason or whether her action can merely be rationalised in the light of some of her reason-states. If we take that distinction into consideration we can see that the first interpretation is either too weak or too strong. The fact that an agent engages in practical reasoning about means to an end does not show that the agent takes herself to have reason in any significant sense, because the agent reasons only about the means to a *given* end. The agent considers only how to achieve the end, rather than the reasons for and against pursuing that end. In that sense the first interpretation is too weak. It is too strong, however, if acting for reasons is associated with action that can be rationalised in the light of some of the agent's mental states and events. Actions that are done for their own sake do not require any kind of practical reasoning, because they are not done *by* doing something else. In that sense requiring means-end reasoning is too strong.

According to a second interpretation, an agent who acts for a reason necessarily treats or endorses some consideration as a reason. There are, *prima facie*, two ways in which an agent's treating as a reason can be construed. Firstly, treating as reason may consist in the agent's having and acting on higher-order beliefs concerning which considerations are—or count as—reasons for actions. Secondly, it may be construed as a genuine and irreducible psychological event or process. Both views, I think, face difficult problems.[72] But I will restrict myself to the point that there is no reason to pursue that proposal in the first place. Consider again Al who goes for a short walk, because he feels like it. There is no reason to assume that Al has a second-order belief concerning which considerations would give him a reason to go for walk, or that Al, in the circumstances, treats or endorses a certain consideration as a reason for going for a walk. Nevertheless, according to the standard-causal model, Al acts for a reason, given that the action is caused in the right way by a reason-state—which is, in that

[72] Compare Wedgwood, *forthcoming*, who argues that reference to higher-order beliefs leads to an infinite regress of ever more higher-order beliefs. Concerning the notion of treating as a reason as irreducible compare chapter 2, p. 89, note 59.

case, an intrinsic desire. Note, again, that I do not deny that treating or endorsing certain considerations as reasons may play an important role in the performance of some actions. Rather, I deny that acting for reasons necessarily involves treating or endorsing something as a reason.

Free Will and the Limits of the Standard-Causal Model

In previous sections we have been concerned first and foremost with the most basic or fundamental aspects of human agency. Agency, as I suggested, should be construed as a hierarchy of kinds of agency, and we might identify the lower boundary of the domain of distinctively *human* agency with the boundary between purposive behaviour and acting for reasons. In this final section I shall say more about the higher, more refined and sophisticated kinds of agency and about the upper boundary of the domain of human agency.

Consider again the proposed hierarchy of kinds of agency: purposive behaviour, acting for reasons, deliberative action, self-controlled action, free and autonomous action. Besides that, moral responsibility is of importance for two reasons. Firstly, many philosophers think that being morally responsible presupposes a certain kind of agency. Secondly, moral responsibility is, of course, itself an important feature of human agency.[73] I am confident that all those kinds and aspects of human agency can be accounted for in standard-causal terms. I cannot argue for this claim in detail, nor do I have the room for detailed accounts of all the mentioned kinds and aspects of agency. But I will show how those tasks can be carried out. All the higher kinds of agency are built on top of—or on the basis of—lower kinds of agency. That means than an account of the higher kinds of human agency can be given by adding further conditions to the account of acting for reasons. What needs to be shown, then, is that those further conditions are compatible with the standard-causal approach.

Higher Kinds of Human Agency

Let us first turn to deliberative action. Discussing Berent Enç's position, we saw that it is possible to provide an account of deliberative action in standard-causal terms. Deliberation concerns, typically, different means to some end, different ends or

[73] Further, one may distinguish moral action. I shall assume, though, that moral action is just a case of acting for reasons—namely, acting for moral reasons.

courses of action, and the consequences thereof. Enç shows that the process of comparing and evaluating different means, ends and consequences can be implemented by standard-causal mechanisms, provided that the mental states and events, which constitute the deliberative process, are causally efficacious in virtue of their contents. That shows that acting on the basis practical reasoning or deliberation is compatible with the standard-causal approach.

Consider next self-controlled action. I suggested that a plausible condition for acting for reasons is that the relevant mental states must be *integrated* in the agent's motivational system, in the sense that they are *responsive* to other reason-states of the agent. A standard-causal account of self-controlled action can take that as a starting point and formulate further and stronger conditions concerning integration and responsiveness. Note first that self-control can be predicated of actions as well as agents. First and foremost, though, it is a feature of agents—a character trait of persons. An agent *lacks* self-control if her actions are *frequently* weak-willed (akratic or incontinent). And an agent is weak-willed with respect to a certain action, if it is performed intentionally against better judgment; that is, the agent intentionally performs the action even though she judges that it would be best or better not to do it (or to do something else).

The standard-causal theory can account for these features by expanding and strengthening the conditions on accordance and responsiveness between desires, beliefs, judgements and intentions. For instance, it may require that in order for an agent *S* to be self-controlled, the tendency of *S* to act on given desires must usually—or frequently—be diminished by *opposing* judgements and beliefs to a degree sufficient to ensure that the agent acts in accordance with the judgements and beliefs rather than the desires.[74] It may require, in other words, *action-guiding* responsiveness to judgements and beliefs.

Let us turn then to free and autonomous action. Like self-control, freedom and autonomy can be attributed to both actions and agents. A minimal and uncontroversial requirement for free action is that the agent is free from constraints like obstruction, interference, coercion, compulsion, and so forth. Accordingly, an agent *S* is free if *S* is

[74] That is compatible with the view that judgements and beliefs can motivate action, but it is not committed to it. One may require that the agent has desires that are in accordance with the judgements and beliefs and that motivate the actions in questions.

frequently—or in some relevant cases, at least—free from constraint. Note, firstly, that the use of the term *free* in that context is entirely unproblematic, since it means nothing more than the absence of constraint. And secondly, there is no reason to think that this kind of freedom is incompatible with the standard-causal model of agency. In order to obtain an account of free *and* autonomous agency on top of all the kinds of agency discussed so far, the standard-causal theory can incorporate either one, or both, of the following two suggestions. Firstly, the theory can incorporate an account of identification in terms of higher-order attitudes.[75] Secondly, the theory can formulate conditions concerning the acquisition and history of the agent's mental attitudes.[76] The latter approach will require that the agent's mental attitudes have been acquired in a normal and unconstrained manner. One may try to give a positive account of what the *normal* acquisition of mental attitudes consists in by appeal to perception, learning, socialisation, and so forth. Alternatively, or in addition, one may require that the agent's mental attitudes have not been induced—through *brainwashing*—or implanted—by *evil scientists*—or otherwise been acquired in a way that bypasses the agent's exercise of her own cognitive and practical skills. Accordingly, an agent *S* is autonomous if, firstly, *S* is capable of deliberative and rational agency; secondly, *S* is a free and self-controlled agent; thirdly, *S* can identify herself with the attitudes that motivate her actions; and, fourthly, if *S* has acquired her mental attitudes in a normal and unconstrained way.

Finally, let us briefly turn to the issue of moral responsibility. Frankfurt's counterfactual intervener example, which has been introduced in the previous section,[77] strongly suggests that moral responsibility does not presuppose that the agent could have done otherwise (in the categorical or unconditional sense).[78] I shall assume that the counterexample is decisive. Given that, there is no reason to think that

[75] Compare, for instance, Dworkin, 1988.
[76] Compare, for instance, Mele, 1995, especially chapter 9.
[77] See pp. 177.
[78] The argument based on that example is, roughly, the following. Arguably, the scenario supports the following two claims. Firstly, Jones performs an action in the actual scenario, but due to the presence of the counterfactual intervener, he could not have done otherwise. Secondly, Jones is morally responsible for that action. Therefore, alternative possibilities are not necessary for moral responsibility. The counterexample and that argument has been challenged in various ways. For discussion see, for instance, Fischer, 1994 and Ekstrom, 2002. The crucial question, it seems to me, is the following. What matters in cases in which we are not sure whether the agent is responsible: whether the agent had alternative possibilities or whether the agent's act was deliberate or intentional? I think the Frankfurt example strongly suggests that what matters is the latter: what matters are properties of the *actual* action, rather than open possibilities.

the feature of being morally responsible—and the kind of agency that it presupposes—is incompatible with the standard-causal account of agency.[79]

In my response to the challenge of disappearing agency I provided an account of acting for reasons and the associated exercise of control in standard-causal terms. Now we can see how and why the higher kinds of human agency can be constructed on the basis of that, and that the accounts of those kinds of agency are compatible with the standard-causal approach. The remaining question is whether the presented hierarchy of kinds of agency is *complete*. If it is complete, we can say that the accounts of the various kinds of human agency taken together are tantamount to an account of human agency. If it is not complete, we are left with an account of most aspects of human agency. Are there higher, more refined and possibly more valuable aspects of human agency? What springs to mind as a missing kind of agency is, of course, acting with free will.

Plural Control and Indeterminism

Free will, I assume, is the ability to choose and to do otherwise. In the first chapter I distinguished between the following two necessary conditions.

> (AP) *Open or Alternative Possibilities*. It is open to the agent to decide and do otherwise than she actually does. In the circumstances, the agent could have decided and done otherwise.
>
> (SDO) *Self-Determination as Origination*. The agent *herself* determines the decision and action. The agent is not only the cause, but the *source* or *origin* of her action.

In the first chapter I assumed that incompatibilism is true. That means, in particular, that AP is incompatible with the truth of determinism, because AP requires a certain kind of indeterminism; namely, indeterminism with respect to certain courses of actions. If Sue *A*-ed after she considered two alternative courses of action, *A*-ing and *B*-ing, then AP is satisfied only if it was not causally determined that Sue *A*-s—only if there was an objective chance that Sue would *B* instead. (It must be undetermined whether Sue *A*-s or *B*-s either at the time she makes the choice or at the time she

[79] Compare Fischer, 1994 and Fischer & Ravizza, 1998, who use Frankfurt-style examples in order to show that guidance control is sufficient for moral responsibility. Guidance control is construed in terms of causal mechanisms, and it is argued that having guidance control is compatible with causal determinism. Given that, it is clear that such an account of moral responsibility is compatible with the standard-causal approach.

performs the action, or at both the time of choice and action—depending on whether free will is associated with choice or action, or with both choice and action.)

So, AP requires that, with respect to past, present and future actions, for an agent to perform an action with free will it must be the case, at the time of choice or action, that more than one course of action is metaphysically open to the agent in the sense explained. But having alternative possibilities in that sense is not sufficient for having libertarian—that is, incompatibilist—free will. Further, it must be the case that the agent has the power or ability to choose and perform any one of the open alternative courses of action. The agent himself must have the power to determine which of the open pathways will become actual. The agent, in other words, must have *control* over which one of the alternative courses of action she performs. Let us call this kind of control *plural* control, as it involves control over a set of more than one alternative courses of action: control over *which* of the open alternatives will become actual.[80]

In the first chapter I pointed out that a indeterministic version of the standard-causal model satisfies AP and that such a theory can account for self-determination construed as non-deviant causation by one's own reason-states. In this chapter I showed that the standard-casual model can also account for a more refined notion of self-determination as autonomy. However, I granted—to proponents of agent-causation—that it fails to account for self-determination *as origination* and that it fails to account for the associated kind of control, which is plural control. Let us now have a closer look at this last claim.

Randolph Clarke says that the kind of control established by the indeterministic standard-causal theory is 'wholly negative: it is just a matter of the absence of any determining cause of the action'.[81] The standard-causal theory 'fails to secure for the agent the exercise of any further positive powers to causally influence which of the alternative courses of events that are open will become actual.'[82] One way of explicating that is to use the model—or analogy—of forking paths. If we think of life

[80] It is common to associate free will with that kind of control. Compare, for instance, Kane who talks about 'plural voluntary control over a set of options' as necessary for acting with free will (Kane, 1998, especially pp. 109-111 and pp. 133-135). Fischer and Ravizza argue that free will requires a '*dual* power', which is 'the power freely to do some act *A*, and the power to do something else instead', and they say that an agent who has that power has regulative control over the set of alternative courses of action (Fischer and Ravizza, 1998, p. 31). Clarke, 2003, says that libertarian free will requires a variety of control 'to causally influence which of the open alternatives will be made actual' (p. 151). Compare, further, Haji, 1998, chapter 1.

[81] Clarke, 2003, p. 96.

[82] Ibid., p. 133.

or the way in which the world unfolds as a path, then there is only one path if determinism is true. However, if indeterminism is true, the paths fork at certain points, and if AP is satisfied, then some of those forking paths represent alternative courses of action open to agents. In order to have free will, however, the paths must not only be open to an agent—it must not only be a matter of chance or probability which path becomes actual. In addition, the agent must have the to power to make one path actual and to 'close off' the others.[83] The indeterministic standard-causal theory establishes only the first, but not the second part of that analogy; it establishes forking paths, but not the required agential power to determine which of the paths will become actual.

Clarke also says that the standard-causal model fails, because the kind of agential control it provides does not go beyond the kind of control established by compatibilist accounts of free will. Consider again Sue's *A*-ing. On the standard-causal model, Sue's exercise of control consists in non-deviant causation by reason-states that are directed towards *A*-ing. That kind of control is compatible with both determinism and indeterminism. What the theory requires is non-deviant *causation*, not non-deviant causal determination. To assume indeterminism does not undermine control, but it does not add anything either. To grant that Sue might instead have *B*-ed for reasons does not show that Sue had the power to determine whether she *A*-s or *B*-s, because whether she *A*-s or *B*-s is a matter of chance or probability (which does not entail that it is random or accidental).

Libertarians are convinced that free will requires not only the falsity of determinism, but also a kind of control that goes beyond a kind of control that is compatible with determinism. I agree that the standard-causal model cannot account for plural control. The theory construes the exercise of agential control as constituted by a causal process, which takes place between an action and a mental state that rationalises, guides and causes *that* action. In cases in which the agent considered more than one option, and decided to do one thing rather than another, we can say that the agent formed an intention to pursue one action *rather than* any of the considered alternatives. But that does not show that the agent had what libertarians mean by

[83] Compare Fischer, 1994, who refers to Borges' story of 'The Garden of Forking Paths' (p. 3). Compare also Haji, 1998, who refers to Feinberg's analogy of 'life as a kind of maze of railroad tracks' (p. 17).

plural control. By forming that intention the agent settled on *one* course of action and the content of the intention guides only the performance of that action.[84] Something similar holds for actions that issue from desires and beliefs. The desire to attain *E* and the belief that one can attain *E* by *A*-ing can rationalise and guide only an *A*-ing. Control as non-deviant causation by reason-states is always directed, so to speak, towards *one* course of action, because reason-states can constitute control only with respect to *the* course of action, which is rationalised by them.[85]

Plural Control and Self-Constitution

Robert Nozick and Robert Kane argued that a crucial element of acting with free will is what I shall call the element of self-constitution.[86] They think that libertarian free will can be captured in standard-causal terms, because self-constitution can be captured in standard-causal terms. I will first introduce the notion of self-constitution, and then I will discuss Robert Kane's theory.[87] I will argue that it fails to account for free will, because it fails to account for plural control.

Consider once more the agent *S* who faces the choice between *A*-ing and *B*-ing. *S* considers the reasons R_A for *A*-ing and reasons R_B for *B*-ing. Assume that *S* decides to *A*, and that this choice is caused and guided by R_A. Assume further that R_A does not causally determine *A*-ing. Rather, R_A renders *A*-ing probable, such that there is an objective chance that S chooses to *B* instead.

Both Nozick and Kane think that acting with free will requires that the action is causally undetermined in the way outlined. If *S* is to act with free will, the reasons, R_A and R_B, must leave it open whether *S A*-s or *B*-s. However, both Nozick and Kane think that the following must be added to a characterisation of the situation. If *S* chooses to *A*, for instance, *S* chooses to *A for* the reasons and thereby *S makes* the reasons in favour of *A*-ing prevail *by acting on them*. Further, making choices of that kind has an effect on some of the agent's psychological and dispositional properties.

---

[84] Compare Mele, 1992, who argues that to form an intention is to settle on one course of action.

[85] That does not mean that an agent never exercises any control over whether to perform one course of action *rather than* another. Assume that Sam forms the judgement that it would be better to *A* rather than *B* before forming an intention to act, and that Sam is disposed to choose and act in accordance with his better judgement. Accordingly, it is then probable that Sam will *A* rather than *B*, and there is a sense in which Sam has control over whether to *A rather than B*. But that kind of control is different from having *plural* control over *A*-ing and *B*-ing.

[86] Nozick, 1981, pp. 294-316 and, Kane, 1998, especially chapter 8.

[87] I will restrict my critique to Kane's position as it encompasses all the important aspect of Nozick's earlier proposal.

Once the agent has chosen to do one course of action for one set of reasons, it will be more likely, in similar future circumstances, that the agent chooses again to act for those reasons. That is to say that the choice changes the agent's disposition how to evaluate the reasons and how to act accordingly. Further, the choice has the *normative* implication that it commits the agent to acting for certain reasons in certain circumstances. In that way choosing and acting with free will has a significant effect on the agent's future actions and on the agent *herself*, as it results in changes with respect to the agent's psychological constitution—that is why I shall call this feature of acting with free will the aspect of *self-constitution*.

Self-constitution is, as far as I can see, perfectly compatible with the standard-causal approach. The problem, though, is that it is difficult to see how it helps with free will, because it is difficult to see how self-constitution is of any relevance to plural control. Let us now have a closer look at Kane's position.

Robert Kane has offered a very insightful, but also very complex account of free will. I have to restrict my discussion to those elements of his account, which I deem central and crucial. Probably *the* most important element is an account of a certain kind of decisions, which Kane calls 'self-forming willings'.[88] These decisions bear all the marks of self-constitution, as introduced above. Kane's preferred example of a self-forming willing is a moral choice by which the agent settles an inner conflict between two courses of action: the agent thinks it is morally required to do one thing, but is tempted to do something else instead. Being undecided, the agent must settle the conflict by making a decision in favour of one of the two alternatives. Generally, a decision is a self-forming willing only if it satisfies all of the following conditions.

(1) The alternative courses of action are genuinely open to the agent.

(2) The agent has reasons for all the alternative courses of action.

(3) No matter what the agent chooses to do, he acts and chooses *for* those reasons.

(4) The agent *makes* those reasons the ones she wants to act on—more than any others—*by* choosing to act on them.[89]

Such choices are self-constitutive—or self-forming, as Kane says—because they 'structure and reorganise the motivational structure' of the agent in a particular way; had the agent decided otherwise, the motivational structure would have been

[88] Compare Kane, 1998, especially pp. 133-142.
[89] Compare ibid., p. 135.

reorganised different correspondingly.[90] The agent wants to do both courses of action and both are reasonable in the light of her beliefs, desires and values. Because of that the agent will identify herself with her choice and action, no matter what she chooses to do. Further, the agent decides which action she prefers *by* making the decision, because 'what [agents] do by choosing is to make one set of reasons prevail over others then and there as motivators of action'.[91]

The question we have to ask is whether the theory offers an account of control that goes beyond causation and guidance by reason-states; in particular, we want to know whether the theory can account for plural control in standard-causal terms. Kane acknowledges that having free will requires that the agent has 'plural voluntary control', as he calls it, and he thinks that his theory can account for it. So far, however, we have not seen anything that show us how standard-causal processes can constitute plural control.

Kane's account of plural voluntary control requires that the agent performs the action on purpose and for reasons, in a sense that encompasses the conditions (1) to (4), and that the agent is not coerced, compelled, or controlled by other agents.[92] But all that does not add anything to a positive account of control. Further, all that is in line with the account of free and autonomous agency that I have outlined at the beginning of this section—an account that does not require plural control. Addressing the sort of scepticism about plural control that I am advocating, Kane says that

> […] it does not follow that because you cannot guarantee which of a set of outcomes occurs beforehand, you do not control *which* of them occurs, *when* it occurs.[93]

Plural control, as Kane points out, has to be distinguished from 'antecedent determining control'. When an agent has antecedent determining control, the agent can 'guarantee or determine which of a set of outcomes is going to occur *before* it occurs'.[94] Plural voluntary control is not like that, since the agent determines the outcome *by* and *when*—and not *before*—making the choice.

Two things must be noticed here. Firstly, that point is only *negative*. It tells us only what plural control is *not*, and in what way it is different from another kind of

[90] Compare ibid., p. 137.
[91] Ibid., p. 135.
[92] Ibid., pp. 142-143.
[93] Ibid., p. 134.
[94] Ibid., p. 144.

control—namely, antecedent determining control. It does not tell us what plural control is in positive terms, and it does not show how and why it can—or cannot—be realised by standard-causal processes. Secondly, the point does not distinguish plural control *uniquely*, because almost exactly the same point can be made with respect to control as non-deviant causation by reason-states. Having that kind of control, the agent does *not* guarantee or determine the outcome *before* performing the action, because the agent exercises control *by* performing an action that is caused, rationalised, guided by and responsive to mental attitudes—not before performing it.

Kane then says that having 'plural voluntary control over a set of options implies being able to bring about *whichever* one (of the options) *you will, when you will* to do so'.[95] Now, that *is* a positive claim. But it is merely conceptual; it characterises what libertarians *mean* when they require plural control. It does not tell us anything about why and how plural control can be understood in standard-causal terms—why and how it can be realised or constituted by standard-causal processes.

Finally, let us return to the case in which the agent settles a moral conflict by making a free choice. Kane uses that case in order to explain the proposed account of plural voluntary control. Commenting on it, he says the following.

> If [agents] fail [to do what they take to be morally required], it will be because they did not *allow* their [moral] efforts to succeed. They chose instead to make their self-interested or present-oriented inclinations prevail […].[96]

This passage is particularly telling. In the attempt to explain how plural control can be constituted by an indeterministic standard-causal process, Kane says that the agent *allows* one set of reasons to prevail. But clearly, this *act* of allowing—or not allowing—certain motives to prevail is part of the very phenomenon that is *in need of explanation*. Either that act of allowing a motive to prevail can itself be accounted for in standard-causal terms, or we must assume that the agent performs it by exercising an agent-causal—or otherwise irreducible—power. Otherwise, the act of allowing a motive to prevail must not be mentioned in an account of plural control. Kane rejects agent-causal together with all other non-reductive theories of agency[97], and he does not provide an account of the act of allowing a motive to prevail in standard-causal

[95] Ibid., p. 134.
[96] Ibid., p. 133.
[97] Compare ibid., p. 116.

terms. Yet Kane cannot avoid alluding to such a notion in order to explain plural voluntary control.

Kane's conditions and remarks concerning plural control are helpful and important. Kane rightly rejects the sceptics who expect plural control to be like 'antecedent determining control'. But given that the offered standard-causal account of control by reasons differs from antecedent determining control as well, that remark is beside the point. What the sceptic should demand is not antecedent determining control, but an *account* of plural control that is *on a par* with the offered standard-causal account of control as causation and guidance by reason-states. In other words, the sceptic should demand an account of how and why a standard-causal process can constitute or realise plural control. Ultimately, Kane has failed to provide that.

Free Will: Why We Don't Have it, and Why That Doesn't Matter

I will now argue that we do not have reason to believe that we have libertarian free will. The argument consists of two main parts. The first part shows that there is no theory of agency available that can account for free will. In the first chapter I showed that non-reductive accounts cannot account for free will, and in this chapter I completed my arguments for the claim that reductive standard-causal theories cannot account for free will either. That completes, in connection with the rejection of volitionism and pluralism, the first part.

In the second part I argue that not having free will is less drastic than libertarians think it is, primarily for the following two reasons. Firstly, Frankfurt-style example give us good reason to think that moral responsibility does not presuppose plural control—hence, that it does not presuppose free will. Secondly, free will concerns an agent's ability to choose between alternative courses of action for reasons. We have seen that the standard-causal theory can account for that ability insofar as it can account for deliberative and autonomous acting for reasons. Further, I will argue that it is not as obvious or straightforward to see what the additional value of having plural control consists in as libertarians think it is. Before turning to the second part, though, I shall say more on open or alternative possibilities.

At the core of the condition of open and alternative possibilities (AP) is the following proposition.

(P) The agent could have done otherwise.

Some philosophers have argued that P admits of a conditional analysis, which is, generally and roughly, of the following form.

(CP) The agent *S would* have done otherwise, if *S* had wanted (decided, intended, or willed) otherwise.

It is common to point out that the antecedent of CP must not mention other *acts* of the agent, since that raises only the question whether the agent could have done otherwise with respect to them. In particular, the conditional analysis must not refer to mental acts such as the agent's willing or making of decisions. Rather, to have at least initial plausibility, the conditional analysis must refer to mental *states* or non-actional mental *events* in the antecedent.[98]

Incompatibilists and libertarians reject that analysis. They insist on what is called an unconditional or categorical interpretation of P. Firstly, even the version of CP that refers only to mental events is subject to counterexamples. Consider Sam who *A*-ed because of an irresistible desire to *A*. It is true that Sam would have done otherwise, had he not had that desire—had he had different reason-states and events. But, intuitively, it is false that Sam could have done otherwise in the circumstance—Sam was not able to do otherwise in the circumstances.[99] Secondly, the conditional analysis, CP, misses the very point of what it is to have or act with free will. Intuitively, to say that an agent acted with free will is to say that that very agent—the same person with exactly the same psychological dispositions, thoughts, desires, beliefs, intentions, and so on—could have done otherwise in exactly the same circumstances at exactly the same time. By making the action counterfactually dependent on the agent's reason-states, the conditional analysis fails to capture the nature of free will, which is highlighted by counterexamples of the kind just presented.

[98] Compare, for instance, Davidson, 1980, essay 4. For discussion see, for instance, Berfosky, 2002, pp. 185-186 and Kane, 1998, pp. 57-59.
[99] See Chisholm, 1964, and Berofsky, 2002, p. 186.

Throughout I assumed that incompatibilism about free will is true, and I have just presented two reasons in support of the claim that CP fails as an analysis of P. The conditional analysis, CP, is usually deployed by compatibilists. But it is subject to counterexamples and I agree with libertarians that it does not capture what acting with free will consists in. However, compatibilism is not committed to CP. That is, the falsity or inadequacy of CP does not entail that compatibilism, in general, is false or inadequate—nor that incompatibilism is true. However, compatibilism in general faces the so-called consequence argument for incompatibilism, which has received a great deal of attention in the more recent literature on free will and which is generally considered to be a very strong and convincing argument.[100] The conclusion of that argument says, in effect, that the ability to do otherwise is not compatible with determinism. Compatibilists, of course, can avoid engaging with that argument by denying that free will is the ability to do otherwise. In response to that I think it is credible to insist, on intuitive grounds, that free will *is* the ability to do otherwise—and that acting with free will requires that the agent could have done otherwise. I shall therefore continue to assume that free will *is* libertarian—that is, incompatibilist—free will.

Where does this leave us with the assessment of the standard-causal theory? To begin with, note that the categorical reading of P admits itself of two interpretations. On a *weak* construal, P requires only a metaphysically open future in the sense that it is not causally determined which of the open alternatives the agent is going to pursue. On a *strong* construal, however, P means that the agent was *able* to do otherwise in the sense that, firstly, the agent had metaphysically open alternatives, and secondly, that the agent had the power to perform either one of the open alternatives—in other words, that the agent had plural control over the alternatives.[101]

---

[100] The most prominent and most detailed formulation can be found in van Inwagen, 1983, who states the argument informally as follows: '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' (p. 16).

[101] Note that this distinction is helpful, but it is not of real importance. The strong construal merely incorporates the insight that, if an agent is *able* to do otherwise, objective alternatives have to be supplemented with plural control—dual power, plural voluntary control, self-determination as origination, or whatever label one prefers. The weak construal, on the other hand, highlights that we can distinguish between having open alternatives and having control over them. To distinguish between the strong and the weak version is therefore, first and foremost, a matter of exposition.

We saw that the standard-causal theory is compatible with the weak construal of the categorical version of P, since it does not require that the causal connections between reason-states and actions are deterministic. That is interesting insofar as it shows that the fact that an action is causally undetermined is compatible with the fact that agent exercises control in performing it; it makes clear, in other words, that the fact that an event is undetermined does not entail that it is random or that it is under no one's control. But the fact that the weak construal of P is compatible with the standard-causal approach is uninteresting insofar as the weak construal is only necessary, but not sufficient for free will. More interesting—and more controversial— is the strong construal, which incorporates the requirement of plural control. I argued that the standard-causal theory cannot account for plural control. Given all that, we can conclude that the standard-causal is compatible with both CP —the conditional analysis of P—and with the weak construal of the categorical interpretation of P. In that sense, it is true that some agents could have done otherwise with respect to some actions, if the standard-causal theory is true. However, the standard-causal theory is incompatible with the strong construal of the categorical reading. In that sense, it is false that some agents could have done otherwise with respect to some of their actions; that is, *in that sense*, no agent could ever have done otherwise, if the standard-causal theory is true.

Not Having Free Will

Let me now explain why that result is less drastic as it seems to be—or why it is less drastic than libertarians think. In a first step, let us recall what is *not* entailed by the assumption that the standard-causal model is correct. Assuming that the standard-causal theory is true, it does not follow that agents do not exercise control over their behaviour. The theory accounts for control as non-deviant causation by—guidance by and responsiveness to—reason-states. Secondly, it does not follow that human agents are not morally responsible for their actions—or that we are not justified in holding others morally responsible for some of their actions and consequences thereof. Frankfurt's counterfactual intervener example strongly suggests that having plural control and libertarian free will is not necessary for being responsible. Thirdly, it does not follow that we cannot distinguish between free and unfree actions. As explained above, the standard-causal approach has resources to distinguish between, and account for, important features of human agency, including rational, deliberative, free

and autonomous agency. Hence, it does not follow that agents cannot govern themselves—that they are not capable of self-determination.[102]

Finally, the standard-causal model is compatible with at least the following two interpretations of P (the claim that, with respect to a certain action, the agent could have done otherwise). Firstly, if a probabilistic version of the theory is true, then it is possible that some agents could have done otherwise in the sense that there was an objective chance that they acted otherwise. Secondly, P requires that agents *could have* done otherwise. The reason for this focus on the past may be that free will has often been treated as a necessary condition for moral responsibility—and usually we hold people morally responsible for their past actions or consequences thereof. But P could easily be reformulated as to apply to past, present, and future actions. What is crucial, according to the categorical reading, is that the circumstances—including the agent's thoughts, dispositions and character traits—are being held constant. The categorical reading focuses on the agent in the *same* circumstances at a particular point in time. The conditional version considers counterfactual scenarios that are *similar* to the actual situation; in particular, scenarios in which the some of the agent's reason-states are different. There is, however, a rather different interpretation of what our ability to do otherwise consists in, which compares the actions of the same agent in *similar* circumstances at *different times*. Frequently, when people say that they can do—or are able to do—otherwise, they mean that they *will* do otherwise in case they find themselves in similar circumstances in the future. By that we mean often that we are able to do *better* in the future, because we have learned the lessons from our past experiences. The ability to learn from our past actions and their consequences is no doubt essential to human agency, moral responsibility and personhood. Clearly, this ability for self-improvement does not presuppose plural control. It is entirely unproblematic, and it is certainly compatible with the standard-causal approach.

[102] *Prima facie*, there are two ways in which the standard-causal theory can account for self-determination without presupposing plural control. Firstly, self-determination can be identified with the kind of free and autonomous agency that has been outlined above (pp. 196). Alternatively, self-determination can be construed being governed by one's *own* reasons: I govern myself insofar as my actions are guided by *my* desires, beliefs and intentions. That approach requires an account of the involved notion of *ownership* in standard-causal terms, which is likely to be tantamount to the suggested account of autonomous agency. (Just consider that an agent's reason-states might be defined as being her *own* if and only if, firstly, the agent can identify herself with them or with being motivated by them, and secondly, the agent has acquired them in a *normal* way—in a way that does not involve deception, brainwashing, coercion etc.)

Given that, we can conclude that the standard-causal model does not entail that agents could not have done otherwise in every sense—or under any interpretation.[103]

The Purported Value of Free Will

That completes the first step of my case for the claim that not having free will is not as drastic as one might think. In a second step I will now argue that our intuitions concerning the value of free will and plural control are less compelling and less straightforward than libertarians claim they are.

Consider the following example.[104] An agent, Alice, faces a choice between two courses of action; between telling the truth or a lie concerning a particular matter. If Alice is able to act with free will, then she is able to tell the truth and she is able to tell a lie, in the sense that she has plural control over the two alternatives. If Alice has free will, then there is a possible world in which Alice tells the truth and a world in which she tells a lie, and those two worlds are identical or alike in all relevant respects up to the point at which Alice decides what to do.

That seems to be a straightforward and unproblematic case of an agent's acting with free will. Libertarians, though, face difficult questions. The first question is *when* or *at what point* in the process, which leads to the action, does Alice exercise free will? Approaching an answer to that question, we have to distinguish between two cases.

Firstly, let us assume that Alice is going through a *proper* deliberative process. Alice weighs the reasons for and against the alternatives, she forms a judgement and an intention on the basis of that evaluation, and she performs an action on the basis of the judgement and in accordance with the intention. Further, we assume that Alice exercises free will in doing so. That means that Alice is at some point in that process able to go *either way*; at some point in that deliberative process the path, which lies ahead of Alice, *forks*. The three most plausible candidates for that point are the formation of the judgement, the formation of the intention and the performance of the action. Libertarians agree that the act of free will is either the act of making a decision

[103] There is, of course, one further interpretation of 'could have done otherwise' that is compatible with the standard-causal model, which construes the ability in question as a general ability of the agent. An agent, for instance, could have lifted that stone, which weighs 20 kilograms, in the sense that the agent is, generally, able to lift a stone that weighs 20 kilograms (as opposed, say, to not being able to lift a stone that weighs 200 kilograms).

[104] I borrow the example from van Inwagen, 2000.

(that is, the act of forming an intention) or the performance of the action. Assuming that Alice's choice is rational, we can assume that the decision is based on the judgement, and that the action is in accordance with the decision.

The problem, of course, is that we have to hold constant the judgement if we want to identify the act of free will with the decision, and we have to hold constant the decision, if we want to identify it with the action. Suppose that in the actual world Alice judges that it is better to the tell truth. She then decides—forms the intention—to tell the truth, and, subsequently, she tells the truth. In the alternative scenario, Alice either decides to tell a lie, even though she judged it is better not to, or she tells a lie, even though she decided not to. In that case, deciding against her judgement or acting against her decision would count as an exercise of libertarian free will. But to anyone who is not concerned with the question of whether or not Alice exercises free will, Alice's action will seem to be a clear instance of weak-willed and irrational behaviour. Weak-willed agents act intentionally against their better judgement or decision, and they do so freely in the sense that their actions are not compelled or forced.

Alice's free choice, it emerges, has lots in common with a weak-willed and irrational choice. In fact, apart from the purported fact that Alice exercises plural control, her choice is indistinguishable form weak-willed and irrational choices. But choices of the latter kind are certainly not the result of the exercise of a particularly refined and valuable form of agency. Arguably, they are not the result of the exercise of a capacity at all; rather, they are the result of a failure to exercise one's ability for rational agency properly.

However, libertarians can object that the first interpretation of Alice's choice does not provide a correct reconstruction of acting with free will. That brings us to the second interpretation, which is in line with Nozick's and Kane's account of free will. According to Nozick and Kane, the agent, whenever she acts with free will, makes one set of reasons prevail by deciding to act on them. On that interpretation, Alice is going through a deliberative process, which differs from the one described. Alice is undecided. She fails to recognise a set of reasons that outweighs the opposing reasons. That is why she must settle the practical problem by deciding which set of reasons to act on. Does that mean that Alice does not form a judgement at all? Either we say that her judgement is identical with the act of choice by which she decides to

act for one set of reasons, or we say that she does not form a judgement at all. The former option, it seems to me, is not very plausible. Arguably, the kind of judgement in question must *compare* the alternative courses of action; it is a judgement concerning which course of action is best, or better than the available alternatives. Further, judgements of that kind must be based on reasons in the sense that the formation the judgment is based on the agent's recognition that she has most reason to pursue one course of action—or good reason to pursue one course of action rather than another. Given that, it would be wrong to say that Alice forms a judgement—failing to recognise that there is reason to pursue one course of action rather than the other, she *cannot* form a judgement. Both the decision to tell the truth and the decision to tell a lie would be based on reasons, but neither decision would be based on a judgement. On that second interpretation, it seems clear *when* or at *what point* Alice acts with free will; namely, when she makes the decision to act on one set of reasons. Identifying the act of free will with a decision of that kind, libertarians can avoid the objection that acting with free will may in some cases be irrational or virtually indistinguishable from weak-willed action. For on that interpretation, actions will typically be in accordance with decisions, and decisions will typically be based on reasons.

Given that interpretation, what does the value of free will consist in? A libertarian might say the following. Alice has reason to tell the truth, and she has reason to tell a lie. She does not think that she has better reasons for either course of action. She cannot judge what is better, and she must, therefore, settle the issue by making a choice. The standard-causal theory can account for choices of that kind. But it cannot account for Alice's having plural control over the alternatives. Human agents, however, can be the *true sources*, *origins* or *authors* of their actions, and they are the true sources, origins or authors of their actions *only if* they have plural control—only if they have libertarian free will. And only as the true sources, origins or authors of their actions, human agents can make a *real* difference. Of course, that is a redescription rather than an explanation of the value of free will. But one should not expect an explanation of the value of free will, since having free will is *intrinsically* valuable. Such redescriptions help us to understand why having free will is intrinsically valuable, but they do not explain its value by referring to an external good—by telling us what free will is *good for*.

I acknowledge the point that having free will may be intrinsically valuable, and I think that the presented considerations concerning authorship and difference-making have some plausibility and force on intuitive grounds. But I do not think that they are conclusive or compelling, for the following three reasons.

Firstly, no theory of agency, I argued, can account for plural control. The reductive standard-causal model can account for agential control in terms of non-deviant causation by reason-states. We can see or understand *why* that is an account of *control*. But neither the reductive standard-causal nor non-reductive models can account for plural control. It remains unexplained how plural control *works*. Given that, it is impossible to *assess* the force of the mentioned libertarian intuitions. Given that there is no account of plural control, it remains unclear what the value of having plural control actually consists in.

Secondly, in the first chapter we saw that it is true, *in some sense*, that according to the standard-causal theory the agent is *not* the *origin* or *source* of her actions. The agent is not the origin or source, in the sense that the causal processes that lead to the agent's actions do not originate within the agent. In that purely metaphysical or extensional sense, the agent is not the origin or source of her actions. But from that it does not follow that the agent is not in control, or that the agent is not governing herself, or that the agent is not the author of her actions. Nor does it follow that the agent does not make a real difference. To the contrary, given all the arguments and considerations of the previous sections and chapters, it would not only be misguided but *false* to say that, according to the standard-causal model, an agent—in our case, Alice—does not govern herself and does not make a real difference.[105]

The third and final point is the following. Libertarians hold that no matter whether Alice tells the truth or a lie, her action will be rational—done for reasons—and, in that sense, it will not be random. The standard-causal theory can accommodate both claims—and it can account for both claims. But there is no reason to think that choices of the kind in question are particularly valuable. For in a particular and narrow sense, Alice's choice *is* random, after all. Alice has reasons for both courses of action, but she does not have reason to choose one rather than the

[105] Libertarians may grant that agents make a difference, on the standard-causal account, but deny that they make a *real* difference. However, I fail to see how one could *argue* for that claim (without begging the question).

other. In order settle the practical question what to do, she has to settle on one course of action *rather than* the other one *for no reason*. In that sense, Alice is merely *picking* one course of action. In the supermarket, for instance, one has to pick that yoghurt, or whatever it is, rather than another one, because one does not have any reason to choose that one rather than another one. There is nothing mysterious about that—but the ability to *pick* is also nothing special or particularly valuable.

Libertarians may object that Alice's choice is like picking only in some respects—and only on an abstract level. The important difference is that Alice's choice is self-constitutive, whereas cases of mere picking are not. That is, whether she tells the truth or not in the present situation will have an influence on choices in similar future situations—by reorganising her motivational structure, as Kane thinks. Given what I said in the section on self-constitution, however, it is clear that this response is beside the point. Firstly, the standard-causal model is compatible with the notion of self-constitution. And secondly, self-constitution does not help with libertarian free will, because it does not establish the required account of plural control. Assume, for the sake of the argument, that Alice's choice is self-constitutive and that it is, in that sense, different from mere picking. Further, let us say that it is valuable to have the ability to make self-constitutive choices. Given all that, nothing follows with respect to the value of having free will, because having free will is distinct from having the ability to make self-constitutive choices. The former requires plural control, whereas the latter does not—and the ability to make self-constitutive choices, as I argued, does not amount to or constitute having plural control.

That shows, I think, that our intuitions concerning the value of free will are not conclusive. It is not clear what the value of having plural control is, since we do not know, exactly, how plural control works. That, of course, concerns only the intrinsic value of free will. Apart from that, free will is often regarded as very important and valuable insofar as it grounds moral responsibility. But, as already mentioned, I think that Frankfurt-style examples show that plural control—and hence libertarian free will—is not necessary for moral responsibility.[106]

[106] Compare p. 197, note 78.

Conclusion

The different abilities that are characteristic of human agency can be arranged in a hierarchy with the most fundamental and basic ability at the bottom, and with more and more sophisticated abilities on top of that. Free will—the ability to do otherwise—could be allocated in that hierarchy in two ways. It is common to regard free will as the highest and most distinct aspect of human agency. On that view, free will would be at the very top of the hierarchy. It would presuppose rational, deliberate, free and autonomous action, but none of these abilities would presuppose free will. On an alternative view, free will is a presupposition of all distinctively human actions. On that view, even acting for reasons presupposes free will, because one acts for reasons only if one chooses with free will to act in the light of reasons.[107]

However, I have argued that a comprehensive account of human agency can be given without free will in either of those two roles—without free will as the highest and most distinct ability and without free will as a presupposition of all human agency. I have argued that we can give an account of all significant abilities and aspects of human agency without presupposing plural control, and I showed that our intuitions concerning the value and importance of free will are far less straightforward than libertarians think. Bringing all the relevant claims together, I can now state my case against free will. The following argument is not supposed to show that free will is impossible, nor that it is impossible that *we* have free will, nor that free will does not exist. Rather, it is an argument for the proposition that we do not have reason to believe in our having free will. It goes as follows. (Recall, once more, that free will is libertarian free will, as I assume that incompatibilism is true.)

In connection with issues concerning reason-explanations of actions and their relation to causal explanations of bodily movements, I argued in the second and third chapter that the standard-causal theory provides the best available account of human agency. Further, I showed that the standard-causal theory does not and cannot account for free will, because it cannot account for plural control.

(1) The best theory of human agency cannot account for our having free will.

The standard-causal approach is a reductive approach. In the first chapter I showed that non-reductive alternatives—in particular, agent-causal theories—cannot account

[107] Compare Searle, 2001.

for free will either. Further, I rejected volitionism and pluralism in the first and the second chapter, respectively. Given that those four positions exhaust the options, proposition (1) can be strengthened as follows.

(2) There is no account of our having free will available.

In a fairly recent article, Peter van Inwagen reaches a similar conclusion.[108] For van Inwagen, however, 'free will undeniably exists'—by which he means, I suppose, that it is undeniable that *we* have free will. According to van Inwagen, free will is, therefore, 'a mystery': we have free will *and* there is a 'strong and unanswered *prima facie* case for its impossibility'.[109]

Given (2), we are faced with the following disjunction: either we believe in a *mystery* or we abandon the belief that we have free will. Obviously, I do not think that free will undeniably exists. One would have good reason to believe in free will, despite its mysteriousness, if one thinks that it is necessary for moral responsibility. We saw, though, that there is good reason to abandon that assumption.

(3) There is good reason to believe that free will is not necessary for moral responsibility.

Many philosophers think having free will is intrinsically valuable. Against that I argued in this section that our intuitions with respect to the intrinsic value of free will are not conclusive, since acting with free will can be virtually indistinguishable from weak-willed action or mere picking. Others think that it is *just obvious* that we have free will. They reject the proposition that we need *reason* to believe in our having free will. It is true that the majority of philosophers writing on free will do not question its existence—or our having it. That is partly because most philosophers deal with the question whether free will is *compatible* with the thesis of causal determinism—they are concerned with its compatibility, rather than its existence. But that does not mean that the existence of—or our having of—free will is uncontroversial, or that it cannot be questioned. To me, and to many others, it is not *just obvious* that we have free will.[110]

---

[108] Van Inwagen, 2000.
[109] Ibid., p. 2.
[110] Note that compatibilists about free will are with me, rather than the libertarian on that matter. Compatibilists claim that we do have free will and that it is compatible with determinism. But the employed conception of free will differs from the libertarian conception, which I have used throughout, in significant respects. Most notably, on the compatibilist conception it is not necessary for an agent to

(4) Our intuitions with respect to the existence and value of free will are inconclusive.

I argued that the standard-causal theory can account for all abilities that are characteristic for human agency, except free will. Further, there is reason to endorse the standard-causal theory, as I showed that it is the best theory available.

(5) The standard-causal theory can account for all aspects of human agency except free will.

(6) There is reason to endorse the standard-causal account of human agency.

If one questioned that we are capable of rational, deliberative, free or autonomous agency, we could, in response, refer to the standard-causal model. That model accounts for all those capacities, and it shows how it is possible that we are both part of the natural order and capable of all the kinds of agency in question. That is, the standard-causal model gives us reason to believe that we are capable of rational, deliberative, free or autonomous agency. Given my arguments, though, it does not give us any reason to believe that we have free will. That completes my case against free will. Taken together, the six claims show that we do not have good—compelling, strong or sufficient—reason to believe in our having the mysterious ability called free will. I shall close with some remarks concerning the overall dialectic.

Firstly, the case against free will concerns the libertarian and incompatibilist conception of free will, according to which having free will presupposes having plural control. I argued against the conditional analysis of alternative possibilities and I referred to the consequence argument against compatibilism, which shows that the ability to do otherwise is incompatible with determinism. As I understand it, free will *is* the ability to do otherwise. Compatibilist theories that deny that we have free will in *that* sense deny, in effect, that we have free will. Those theories may well provide interesting and viable account of related kinds of agency, such as deliberate and autonomous agency, but we should not say that they account for our having free will.

Secondly, the defended standard-causal model of agency is compatible with both causal determinism and causal indeterminism. It construes agential control in terms of non-deviant causation by reason-states and, in contrast to some compatibilist theories of free will, it does not require that those causal processes are deterministic. From

---

have plural control in order to have free will. Compare, for instance, Dennett, 1984 and Frankfurt, 1971, reprinted in Frankfurt, 1988, as essay 2.

that, though, that it does not follow that deliberative, free and autonomous agency is compatible with both determinism and indeterminism. One may argue, for instance, that free and autonomous agency requires undetermined events at some stage in the causal processes that result in free and autonomous actions.

A final remark concerns the revision of common sense intuitions. Whether or not a theory of human agency is *plausible* depends to a large extend on whether it reflects—or is in accordance with—our intuitions concerning agency. The proposed view is in line with most of our intuitions. Nevertheless, opponents may insist that it is implausible, since it fails to capture one very important and central intuition; namely, the intuition that we have free will. However, even if there was overwhelming intuitive agreement that we have libertarian free will, it would *only* be a majority view. Further, it must be granted that theories can be *revisionary*. Even if a philosophical theory of human agency must be largely supported by our intuitions, one should not require or expect that it is line with all intuitions. The mere fact that a theory is in conflict with *some* common sense intuitions should not render it implausible. Rather, one must ask further questions such as how many intuitions are violated by a given theory (in comparison to the number of intuitions that support it)? How central or important are those intuitions? I argued that libertarian free will is not indispensable, because it is not necessary for moral responsibility and because our intuitions concerning its existence and value are inconclusive. The proposed position certainly is revisionary for those who think it is *undeniable* that we have *libertarian* free will. But that alone does not provide sufficient reason to discard it.[111]

[111] To be sure, my conclusion and that final remark concern the use of the term 'free will' in philosophy and the considered intuitions of philosophers; in particular, my conclusion and that final remark concern a notion of free will according to which having free will presupposes having plural control.

Conclusion

I defended and argued for the reductive standard-causal model of agency. In the first chapter, I rejected the non-reductive approach and volitionism. I introduced pluralism, and I showed that apparent alternatives, such as emergentism and a Kantian psychology, fall under the introduced positions. In the second chapter, I turned to causalism about reason-explanations. I considered different interpretations of the claim that reasons cause and causally explain actions, and I rejected alternative non-causal theories of reason-explanation. Non-causalism, firstly, fails to account for the metaphysical aspect of acting for reasons, and secondly, it is committed to pluralism. I argued that the pluralistic stance is unsatisfactory, because it does not recognise the fact that the relationship between reasons for actions and the causes of bodily movements is in need of explanation. In the third chapter, I argued that causalism can account for that relationship—which is to say that it can solve what I called the coincidence problem. I showed, further, that a solution to that problem neither requires the assumption that mental types are identical with non-mental types, nor that psychology is reducible. I defended this non-reductive position against the causal exclusion argument. In the fourth chapter, I presented a response to the challenge of disappearing agency by providing an account of agential control, construed as non-deviant causation by reason-states. Further, I argued that acting for reasons neither requires deliberation nor that the agent actively treats some consideration as a reason. In the last section of that last chapter, we saw that the standard-causal model cannot account for libertarian free will, and I argued that we do not have reason to believe in our having free will, under the assumption that incompatibilism is true.

This thesis does not claim to defend an original position in the theory of action. The standard-causal theory is the mainstream position in the analytical philosophy of action and mind. I explained in the introduction why it is, nevertheless, in need of defence. The main contribution of this thesis is that it provides an overall or global argument for the standard-causal approach, and that it responds to global rather than local challenges. The rival agent-causal theory is usually rejected because of problems with the presupposed notion of substance-causation. I assumed the possibility of causation by substances, and I argued that the agent-causal view fails as a theory of agency, as it fails to account for agential control. I showed, further, that this argument

applies to the non-reductive approach in general. The second and the third chapter present my global argument for the standard-causal approach, and in the third chapter I suggest a novel solution to the coincidence problem. In the final chapter, I argued that the challenge of disappearing agency is merely a challenge, not a philosophical problem. Further, I argued that there is no related metaphysical problem of accounting for the agent's role in 'full-blown' agency or 'action *par excellence*'. In my response to the challenge of disappearing agency, I suggested a straightforward solution to the problem of deviant causal chains, which has been overlooked by most philosophers who have attempted to solve this troublesome problem. And in the last chapter, I proposed a non-standard stance with respect to the problem of free will. I assumed that free will, construed as the ability to do otherwise, is incompatible with the thesis of causal determinism. I argued that having libertarian free will presupposes plural control, and I showed that neither the reductive nor the non-reductive model can account for that kind of control. Further, I assumed that being morally responsible does not presuppose acting with libertarian free will, and I argued that our intuitions concerning the value of free will are inconclusive. On the basis of that, I concluded that we do not have reason to believe in our having libertarian free will.

What is distinctive about that approach to the problem of free will is that it shifts the focus from the question of whether having free will is compatible with determinism to the question of whether there is reason to believe that we have plural control and the question of what the value of libertarian free will consists in. The motive behind that shift is, firstly, that there is reason to endorse the standard-causal model, and secondly, that this model is compatible with *both* causal determinism and indeterminism.

The promoted position on free will does not fall under any of the traditional views, such as compatibilism or hard determinism, and I am not aware of any other established name for it either. But that, of course, is not to say that it is entirely original, as it shares some important assumptions and claims with the position known as semi-compatibilism.[1] The defended position, however, does not fall under semi-compatibilism. One important difference is that my position is not committed to the

[1] Compare Fischer, 1994, and Fischer and Ravizza, 1998: 'moral responsibility is compatible with causal determinism, even if causal determinism is incompatible with freedom to do otherwise' (Fischer and Ravizza, p. 53).

claim that moral responsibility is compatible with causal determinism. Rather, it is committed to the related but different claim that moral responsibility does not presuppose plural control—and that it does, therefore, not presuppose libertarian free will. Further, semi-compatibilism concerns first and foremost moral responsibility. Concerning free will it says merely that incompatibilism might well be true; it neither claims that free will is incompatible with causal determinism, nor does it assert or deny that we have free will.

Finally, I shall highlight some of the background assumptions and some remaining issues and problems. I assumed that there are substances—things and beings that persist through change and that possess causal powers—and properties. I assumed that some properties are causally relevant, and I remained neutral on the question of whether events are particulars or instantiations of properties. In the last chapter, I assumed that the ability to do otherwise is incompatible with causal determinism and that Frankfurt-style examples show that moral responsibility does not require plural control.

In the first and second chapter, I motivated but did not fully justify the assumption that actions are events. In the second and third chapter, I responded to different challenges to the claim that reason-states cause and causally explain actions. I presented different ways in which the causal explanatory force of reason-explanations can be construed—by appeal to counterfactuals, intentional *ceteris paribus* laws or underlying mechanisms—, but I did not develop and defend any position in particular.

The third chapter contributes to a defence of non-reductive physicalism, but is by no means a full defence. The proposed solution to the coincidence problem has been presented in rather abstract and broad terms. A full defence would require a more detailed exposition of the proposed strategy. Concerning the causal exclusion argument, I tried to deflate the charge that the proposed position does not show how mental states and event can be causally efficacious independently of and in addition to physical states and events, and I responded to the objection that it establishes merely the causal explanatory relevance of the mental, but not its causal efficacy. It is more than likely, though, that my responses will not convince all opponents. However, I think I did enough in order to show, firstly, that the causal exclusion

argument is not as straightforward and not as pervasive as many think, and secondly, that non-reductive physicalism is a plausible and viable position.

Further, a full account of the standard-causal model of agency would need to say more on the relationship between acting for reasons and acting for good or normative reasons. I suggested using the term 'acting for reasons' in a minimal sense that does not presuppose acting for good reasons, deliberation, means-end reasoning, nor that the agent actively treats something as a reason. It is important to note that on my view the usage of the term 'acting for reasons' is terminological. It is a plausible use, but I do not claim that it is based on *the* correct analysis of the terms—on my view there is no single one correct analysis, as I think that the term is ambiguous. A fully developed theory, though, would need to specify more in detail what acting for good reasons and an agent's treating something as a reason consists in, and how those features relate to acting for reasons in the minimal sense.

Finally, there are some open questions concerning the normativity of rational action and the issue of rule-following. In the fourth chapter, I explained how the standard-causal theory can accommodate the intuition that a reason for action is a rational cause of action. Reason-states do not merely cause and rationalise action, but they cause them in virtue of rationalising them, because an action that is caused by a reason-state in the right—that is, non-deviant—way is a response to the content of a mental state that rationalises the action's performance. One may object, though, that rational action must be construed in normative rather than causal terms, because an agent who acts rationally does not simply manifest a psychological regularity. Rather, an agent who acts rationally *follows* a rule—a norm or standard of rationality. It is not clear to me whether such considerations constitute an objection or a problem. But even if that worry is misguided, it would be interesting to see why. A full defence of the standard-causal theory might therefore include a response to the just outlined point concerning rule-following.

References

Anscombe, G. E. M. (1957) *Intention*, Cambridge: Harvard University Press.

Antony, Louise (1989) 'Anomalous Monism and the Problem of Explanatory Force', *The Philosophical Review*, Vol. 98, 153-187.

—(1995) 'Law and Order in Psychology', *Philosophical Perspectives*, 9, 429-446.

—(1991) 'The Causal Relevance of the Mental: More on the Mattering of Minds', *Mind & Language*, Vol. 6, No. 4, 295-327.

Antony, L. & Levine J. (1997) 'Reduction with Autonomy', *Philosophical Perspectives*, 11, 83-106.

Aristotle (1915) *The Works of Aristotle*, London: Oxford University Press.

Audi, Robert (1997) 'Acting for Reasons', in Mele (ed.) *The Philosophy of Action*, Oxford: Oxford University Press, 75-105.

—(1993a) *Action, Intention, and Reason*, London: Cornell University Press.

—(1993b) 'Mental Causation: Sustaining and Dynamic', in Heil and Mele (ed.) *Mental Causation*, Oxford: Clarendon Press.

Beckermann, Ansgar (1992) 'Introduction – Reductive and Nonreductive Physicalism', in Beckermann et al (ed.) *Emergence or Reduction?*, New York: Walter de Gruyter.

Berofsky, Bernard (2002) 'Ifs, Cans, and Free Will: The Issues', in Robert Kane (ed.) *Oxford Handbook on Free Will*, Oxford: Oxford University Press, 181-201.

Bishop, John (1989) *Natural Agency. An Essay on The Causal Theory of Action*, Cambridge: Cambridge University Press.

Block, Ned (1978) 'Troubles with Functionalism', in Block (ed.) (1980) *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge: Harvard University Press.

—(1997) 'Anti-Reductionism Slaps Back', *Philosophical Perspectives*, 11, 107-132.

Botterill G. & Carruthers P. (1999) *The Philosophy of Psychology*, Cambridge: CUP.

Brand, M. (1984) *Intending and Acting*, Cambridge: MIT Press.

Bratman, M. E. (1987) *Intention, Plans, and Practical Reason*, Cambridge: Harvard University Press.

—(2001) 'Two Problems About Human Agency', *Proceedings of the Aristotelian Society*, 2000-2001, 309- 326.

—(2000) 'Valuing and the Will', in *Philosophical Perspectives*, 14, 249-265

Braun, David (1995) 'Causally Relevant Properties', *Philosophical Perspectives*, 9, 447-475.

Brewer, Bill (1995) 'Mental Causation', *Proceedings of the Aristotelian Society*, Suppl. Vol. 69, 237-253.

Burge, Tyler (1979) 'Individualism and the Mental', *Midwest Studies in Philosophy,* Vol. IV.

Campbell, John (1994) *Past, Space and Self*, Cambridge: MIT Press.

Child, William (1994) *Causality, Interpretation and the Mind*, Oxford: OUP.

Chisholm, Roderick (1977) 'Comments and Replies', *Philosophia*, Vol. 7, No. 1, 597-636.

—(1966) 'Freedom and Action', in Lehrer (ed.) *Freedom and Determinism*, New York: Random House, 11-44.

—(1964) 'Human Freedom and the Self', reprinted in Gary Watson (ed.) *Free Will*, Oxford: Oxford University Press, 2003, 26-37.

—(1975) 'The Agent as Cause', in Brand & Walton (eds.) *Action Theory*, Boston: D. Reidel Publishing Company, 199-212.

Clarke, Randolph (2003) *Libertarian Accounts of Free Will*, Oxford: Oxford University Press.

—(1993) 'Toward a Credible Agent-Causal Account of Free Will', reprinted in Gary Watson, ed. (2003) *Free Will*, Oxford: Oxford University Press, 257-284.

Crane, Tim (1995) 'The Mental Causation Debate', *Proceedings of the Aristotelian Society*, Supplementary Volume LXIX, 211-236.

Cussins, Adrian (1992) 'The Limitations of Pluralism', in Charles & Lennon (eds.) *Reduction, Explanation and Realism*, Oxford: Oxford University Press, 179-224.

David, Marian (1997) 'Kim's Functionalism', *Philosophical Perspectives*, 11, 133-148.

Davidson, Donald (1980) *Essays on Action and Events*, Oxford: Oxford University Press.

—(1993) 'Thinking Causes', in Heil and Mele (ed.) *Mental Causation*, Oxford: OUP.

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

—(2004) 'Two Ways of Explaining Actions', *Royal Institute of Philosophy Supplement*, 55, 25-42.

Danto, A. C. (1965) 'Basic Actions', *American Philosophical Quarterly*, 2, 141-148.

Dennett, Daniel. (1984) *Elbow Room. The Varieties of Free Will Worth Wanting*, Cambridge: MIT Press.

—(1991) 'Real Patterns', *The Journal of Philosophy*, 88, 1, 27-51.

—(1989) *The Intentional Stance*, Cambridge: MIT Press.

Dupré, John (2001) *Human Nature and the Limits of Science*, Oxford: Clarendon Press.

—(1993) *The Disunity of Science*, Cambridge: Harvard University Press.

Dretske, Fred (1988) *Explaining Behaviour. Reasons in a World of Causes*, Cambridge: MIT Press

Dworkin, G. (1988) *The Theory and Practice of Autonomy*, New York: Cambridge University Press.

Ekstrom, Laura (2002) 'Libertarianism and Frankfurt-Style Examples', in Kane (ed.) *The Oxford Handbook of Free Will*, Oxford: Oxford University Press.

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

Fischer, John M. (1994) *The Metaphysics of Free Will*, Oxford: Blackwell.

Fischer, J. M., and Ravizza, M. (1998) *Responsibility and Control. A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Flanagan, Owen (1992) *Consciousness Reconsidered*, Cambridge: MIT Press.

Fodor, J. A. (1991) 'Fodor's Guide to Mental Representation: the Intelligent Auntie's Vade-Mecum', in Greenwood (ed.) *The Future of Folk Psychology*, Cambridge: CUP, 22-50.

—(1989) 'Making Mind Matter More', *Philosophical Topics*, 17, 1, 59-79.

—(1974) 'Special Sciences', *Synthese*, 28, 97-115.

—(1991) 'You Can Fool Some of The People All of The Time, Everything Else Being Equal; Hedged Laws and Psychological Explanation', *Mind*, 100, 19-33.

Frankfurt, Harry (1988) *The Importance of What We Care About*, Cambridge: Cambridge University Press.

Ginet, Carl (2001) 'Reasons Explanations of Action: Causalist versus Noncausalist Accounts', in Robert Kane (ed.) *Oxford Handbook on Free Will*, Oxford: Oxford University Press, 386-405.

—(1990) *On Action*, Cambridge: Cambridge University Press.

Goldman, Alvin (1970) *A Theory of Human Action*, New Jersey: Prentice-Hall.

Hasker, William (1999) *The Emergent Self*, London: Cornell University Press.

Haji, Ishtiyaque (1998) *Moral Appraisability*, Oxford: Oxford University Press.

Hempel, C. and Oppenheim, P. (1948) 'Studies in the Logic of Explanation', *Philosophy of Science*, 15, 135-175.

Horgan, Terrence (1993) 'From Supervenience to Superdupervenience: Meeting the Demands of the Material World', Mind, 102, 555-586.

Horgan, T. & Tye M. (1988) 'Against the Token Identity Theory', in Lepore & McLaughlin (ed.) *Action and Events*, Oxford: Blackwell.

Horgan, T. & Woodward, J. (1985) 'Folk Psychology is Here to Stay', *The Philosophical Review*, 94, 2, 197-225.

Hornsby, Jennifer (1980) *Actions,* London: Routledge & Kegan Paul.

Hume, David (1748) *An Enquiry Concerning Human Understanding*, Oxford: Clarendon Press, 1974.

—(1739) *A Treatise of Human Nature*, Oxford: Clarendon Press, 1975

Jackson, F. and Petit, P. (1988) 'Functionalism and Broad Content', *Mind*, 97, 381-400.

Kane, Robert (1998) The Significance of Free Will, Oxford: Oxford University Press.

—(2002) 'Introduction: The Contours of the Contemporary Free Will Debate', in Kane (ed.) *The Oxford Handbook of Free Will*, Oxford: Oxford University Press.

Kant, Immanuel (1997) *Groundwork of the Metaphysics of Morals*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press.

Kim, Jaegwon (1997) 'Mechanism, Purpose, and Explanatory Exclusion', in Mele (ed.) *The Philosophy of Action*, Oxford: Oxford University Press, 256-282.

—(2000) *Mind in Physical World: An Essay on the Mind-Body Problem and Mental Causation*, Cambridge: MIT Press.

—(1993) *Supervenience and Mind. Selected Philosophical Essays*. Cambridge: Cambridge Studies in Philosophy.

Korsgaard, Christine (1996a) 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations', in her *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press, 188-221.

—(1999) 'Self-Constitution in the Ethics of Plato and Kant', *The Journal of Ethics*, 3, 1-29.

—(1996b) *The Sources of Normativity,* Cambridge: Cambridge University Press.

Kripke, Saul (1973) *Naming and Necessity*, Cambridge: Harvard University Press.

Lennon, Kathleen (1990) *Explaining Human Action*, London: Duckworth.

Lewis, David (1986) 'Causation', in his *Philosophical Papers 2*, Oxford: Oxford University Press, pp. 159-213.

—(1980) 'Mad Pain and Martian Pain', in Block (ed.) (1980) *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge: Harvard University Press.

Lowe, E. J. (2003a) 'Physical Causal Closure and the Invisibility of Mental Causation', in Walter & Heckmann (eds.) *Physicalism and Mental Causation*, Exeter: Imprint Academic, 137-154.

—(1996) *Subjects of Experience*, Cambridge: Cambridge University Press.

—(2003b) 'Substance Causation, Persons, and Free Will', in Kanzian, Quitterer & Runggaldier (eds.) *Persons: An Interdisciplinary Approach*, Vienna, 76-88.

—(1993) 'The Causal Autonomy of the Mental', *Mind*, Vol. 102, No. 408, 629-644.

Macdonald, Graham (1992) 'The Nature of Naturalism', *Proceedings of the Aristotelian Society*, Supplementary Vol. 66, 225-244.

McCann, H. J. (1998) *The Works of Agency*, London: Cornell University Press.

Melden, A. I. (1961) *Free Action*, London: Routledge & Kegan Paul.

Mele, Alfred R. (1997) 'Agency and Mental Action', *Philosophical Perspectives*, 11, 231-249.

—(1995) *Autonomous Agents. From Self-Control to Autonomy*, Oxford: Oxford University Press.

—(2003) *Motivation and Agency*. Oxford: Oxford University Press.

—(1992) *Springs of Action. Understanding Intentional Behaviour*, Oxford: Oxford

Mele A., and Moser K. (1997) 'Intentional Action', in Mele (ed.) *The Philosophy of Action*, Oxford: Oxford University Press, 223-255.

Menzies, Peter (2001) 'The Causal Efficacy Of Mental States', [online], retrieved 17/12/05, http://www.phil.mq.edu.au/staff/pmenzies/OnlinePapers.html

Merricks, Trenton (2001) *Objects and Persons*, Oxford: Clarendon Press

Mill, J. S. (1846) *A System of Logic*, New York: Harper & Brothers.

Moya, Carlos (1990) *The Philosophy of Action*, Oxford: Polity Press.

Nagel, Ernest (1961) *The Structure of Science*, London: Routledge & Kegan Paul.

Nagel, Thomas (1986) *The View from Nowhere*, New York: Oxford University Press.

Noordhof, Paul (1999) 'Causation by Content?', *Mind and Language*, 14, 3, 291-320.

Nozick, Robert (1981) *Philosophical Explanations*, Cambridge: Harvard University Press.

O'Connor, Timothy (1995) 'Agent Causation', reprinted in Gary Watson, ed. (2003) *Free Will*, Oxford: Oxford University Press, 257-284.

—(2000a) 'Causality, Mind and Free Will', *Philosophical Perspectives*, 14, 105-117.

—(2002) 'Dualist and Agent-Causal Theories', in Robert Kane, ed. (2001) *Oxford Handbook on Free Will*, Oxford: Oxford University Press, 337-355.

—(2000b) *Persons and Causes. The Metaphysics of Free Will*, Oxford: Oxford University Press.

Olson, E. T. (1997) *The Human Animal*, Oxford: Oxford University Press.

Owens, David (1989) 'Levels of Explanation', *Mind*, 98, 59-79.

Papineau, David (1992) 'Irreducibility and Teleology', in Charles & Lennon (eds.) *Reduction, Explanation and Realism*, Oxford: Oxford University Press, 45-68.

Parfit, Derek (1984) *Reasons and Persons*, Oxford: Oxford University Press.

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

Pettit, Philip (1992) 'The Nature of Naturalism', *Proceedings of the Aristotelian Society*, Supplementary Vol. 66, 245-266.

Pietroski, P. M. (2000) *Causing Actions*, Oxford: OUP.

Polger, Thomas W. (2004) *Natural Minds*, Cambridge: MIT Press.

Putnam, Hilary (1975) 'The Meaning of "Meaning"', in his *Mind, Language and Reality*, Cambridge: Cambridge University Press.

—(1967) 'The Nature of Mental States', in Chalmers (ed.) (2002) *The Philosophy of Mind*, Oxford: Oxford University Press, 73-79.

—(1999) *The Threefold Cord: Mind, Body and World*, New York: Columbia University Press.

Ruben, D. (2003) *Action and its Explanation*, Oxford: OUP.

Ryle, Gilbert (1949) *The Concept of Mind*, New York: Barnes & Noble.

Rorty, Richard (1979) *Philosophy and the Mirror of Nature*, Princeton: Princeton University Press.

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

Schiffer, Stephen (1991) 'Ceteris Paribus Laws', *Mind*, 100, 1-17.

—(1987) *Remnants of Meaning*, Cambridge: MIT Press.

Schueler, G. F. (2003) *Reasons and Purposes. Human Rationality and the Teleological Explanation of Action*, Oxford: Clarendon Press.

Searle, John (1983) *Intentionality*, Cambridge: Cambridge University Press.

—(2001) *Rationality in Action*, Cambridge: MIT Press.

—(1992) *Rediscovery of the Mind*, Cambridge: MIT Press.

Segal, G. & Sober, E. (1991) 'The Causal Efficacy of Content', *Philosophical Studies*, 63, 1-30.

Sehon, Scott R. (2000) 'An Argument Against the Causal Theory of Action Explanation', *Philosophy and Phenomenological Research*, Vol. LX, 1, 67-85.

—(1994) 'Teleology and the Nature of Mental States', *American Philosophical Quarterly*, 31, 1, 63-72.

Shoemaker, Sidney (1981) 'Some Varieties of Functionalism', *Philosophical Topics*, Vol. 12, 1, 83-118.

Skorupski, John (1999) 'Reasons and Reason', in his *Ethical Explorations*, Oxford: Oxford University Press.

Smith, Michael (1994) *The Moral Problem*, Oxford: Blackwell.

Smith, Peter (1992) 'Modest Reductions and the Unity of Science', in Charles & Lennon (eds.) *Reduction, Explanation and Realism*, Oxford: Oxford University Press, 19-44.

Sosa, Ernest (1984) 'Mind-Body Interaction and Supervenient Causation', *Midwest Studies in Philosophy*, 9, 271-281.

Stoutland, Frederick (1998a) 'Intentionalists and Davidson on Rational Explanation', in Meggle et al (ed.) *Actions, Norms, Values*, 191-208.

—(1986) 'Reasons, Causes and Intentional Explanation', *Analyse & Kritik*, vol. 8, p. 28-55.

—(1976) 'The Causation of Behaviour', *Acta Philosophia Fennica*, 28, 286-326.

—(1998b) 'The Real Reasons', in Bransen and Cuypers (eds.) *Human Action, Deliberation and Causation*, Dortrecht: Kluwer, 43-66.

Strawson, Galen (1986) *Freedom and Belief*, Oxford: Clarendon Press.

Strawson, P. F. (1985) 'Causality and Explanation', in Vermazen and Hintikka (eds.), *Essays on Davidson: Actions and Events*, Oxford: Clarendon Press, pp. 115-36.

Taylor, Richard (1966) *Action and Purpose*, New York: Prentice Hall.

Thalberg, Irving (1984) 'Do Our Intentions Cause Our Intentional Actions?', *American Philosophical Quarterly*, 21, 249-60.

Van Inwagen, Peter (1977) 'A Definition of Chisholm's Notion of Immanent Causation', *Philosophia*, Vol. 7, No. 1, 567-581.

—(1983) *An Essay on Free Will*, Oxford: Clarendon Press.

—(2000) ‘Free Will Remains a Mystery’, in *Philosophical Perspectives*, 14, 1-19.

Von Wright, G. H. (1971) *Causality and Determinism*, New York: Columbia University Press

Velleman, David (2000) ‘What Happens When Someone Acts?’ in his *The Possibility of Practical Reason*, Oxford: Oxford University Press, 123-143.

Watson, Gary (1972) ‘Free Agency’, *Journal of Philosophy*, 72, 205-220.

Wedgwood, Ralph (*forthcoming*) ‘The Normative Force of Reasoning’, *Nous*.

Williams, Bernard (1981) *Moral Luck*, Cambridge: Cambridge University Press.

Wilson, G. M. (1989) *The Intentionality of Human Behaviour*, Stanford: Sanford University Press.

Wittgenstein, Ludwig (1953) *Philosophical Investigations*, New York: Macmillan.

Yablo, Stephen (1992) ‘Mental Causation’, in Chalmers (2002) *Philosophy of Mind*, Oxford: OUP, p. 179-196.

Yaffe, Gideon (2000) *Liberty Worth the Name: Locke on Free Agency*, Princeton: Princeton University Press.