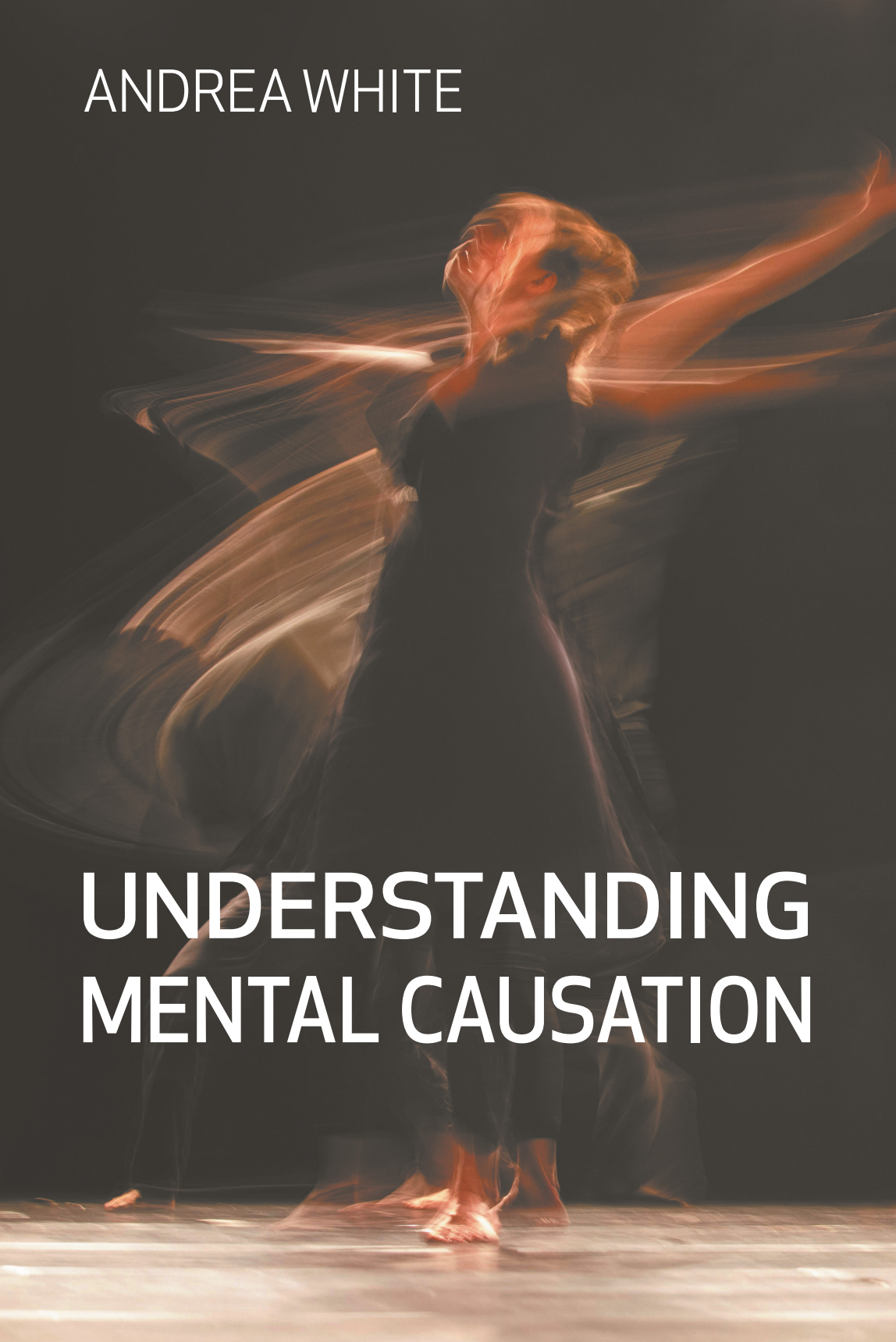
ANDREA WHITE



UNDERSTANDING MENTAL CAUSATION

Understanding Mental Causation

Andrea White

WHITE ROSE
UNIVERSITY PRESS
Universities of Leeds, Sheffield & York

Published by
White Rose University Press
(Universities of Leeds, Sheffield and York)
University of York,
Heslington, York, UK, YO10 5DD
https://universitypress.whiterose.ac.uk

Understanding Mental Causation

First published 2024

Cover Illustration: Photo by Ahmad Odeh on Unsplash
Cover designed by: Kate Petherbridge, WRUP

Print and digital versions typeset by Siliconchips Services Ltd.

ISBN (Paperback): 978-1-912482-52-8
ISBN (PDF): 978-1-912482-53-5
ISBN (EPUB): 978-1-912482-54-2
ISBN (MOBI): 978-1-912482-55-9

DOI: https://doi.org/10.22599/White

Example citation: Andrea White, *Understanding Mental Causation* (York: White Rose University Press, 2024). DOI: https://doi.org/10.22599/White. CC BY-NC 4.0, https://creativecommons.org/licenses/by-nc/4.0/

To access this work freely online via the White Rose University Press website, please scan this QR code or visit https://doi.org/10.22599/White.

Acknowledgements

I am incredibly grateful to the University of Leeds for awarding me the inaugural Leeds Early Career Publishing Prize, which made it possible for me to write and publish this book. It was a wonderful surprise to win the prize and I feel very privileged to be able to share my research with a wider audience.

Most of the research for this book was done during my PhD, and many of the ideas presented here are based on ideas in my PhD thesis. I am very grateful to the University of Leeds for awarding me the University of Leeds 110 Anniversary Scholarship, which made it possible for me to complete my PhD.

And I owe heartfelt thanks to Helen Steward for supervising my PhD and for mentoring me while I wrote this monograph. I enjoyed my PhD immensely and that was in no small part down to Helen, whose thoughtfulness, precision and insightfulness helped me claw my way out of many philosophical puzzles!

I also wish to thank Jennifer Hornsby and Robin Le Poidevin for examining my PhD thesis. Their questions and comments during my viva were inspiring and encouraging.

Many other people have helped me learn, improved my ideas and offered much-needed support or encouragement, including Bill Brewer, William Child, David Hillel Ruben, Maria Alvarez, John Hyman, Heather Logue, Ursula Coope, Johanna Roessler, Hans-Johann Glock and the graduate students of King's College London and Leeds philosophy departments.

I wish to thank the four anonymous reviewers who provided invaluable comments on both the proposal for this monograph and the first draft.

I am very grateful to Kate Petherbridge and Lucy Cook and the team at WRUP for guiding me through the peer review and publication process and for supporting me during what was sometimes a difficult process! I also want to offer deep thanks to Tom Stoneham, whose guidance helped determine the structure of this book and greatly improved it.

I wrote most of this book while working full time as a primary school teacher. I am grateful to all the children I taught for inspiring me and for being their unique wonderful selves. Writing philosophy while working outside academia was certainly not easy. It helped me understand the value and importance of open access publishing. I am very glad that this monograph will be freely available to any interested scholar and wholeheartedly support the growing movement towards open scholarship.

Finally, I want to thank Dan Hinge. You have probably read this book more times than anyone, even me. Your comments and suggestions were extremely helpful and important, and your editing improved the clarity of my writing immeasurably. Thank you for your support and encouragement and above all your patience.

Contents

Introduction

It seems undeniable that our mental life makes a difference, sometimes a big difference, to our bodily life. What we think, what we believe, what we want, what we feel affects what we do with our bodies. I add salt to the sauce because I think that will make it taste better, I water my plants because I want them to grow, I take off my shoes because my feet feel sore, I wince because I remember an embarrassing mistake, I speak hesitantly because I feel nervous. Ordinary experience seems to suggest that what we do with our bodies causally depends, somehow, on what's going on in our minds.

How to understand the causal aspect of the mind–body connection is the subject of this book. Many philosophers have thought that our ordinary experience shows that there is causal interaction between mind and body, or that changes in one cause changes in the other. However, problems start to arise when we try to understand how this could be, given certain assumptions about the nature of reality. For example, suppose you thought, as Descartes did, that the mind is not a material thing. Instead, it is the immaterial part of ourselves that thinks and which is joined with our body but nevertheless distinct from it. If you also assumed that causal interaction could only occur between material things, perhaps because you thought all causal interaction required some kind of physical contact, then it becomes hard to see how mind–body causal interaction is possible. How can the mind have causal effects in the material world if it is not itself material?[1]

In contemporary philosophy of mind, putting together a plausible account of the mind–body connection remains a significant challenge. Philosophers of mind strive to give an account of what the mind is that allows mentality to have causal relevance but which also fits with the most plausible views of what causation in the actual world must be like. This is the problem of mental causation.

[1] This is the most famous objection levelled at Descartes's dualist metaphysics. See Shapiro (2007: 62) for Princess Elisabeth's version of this objection.

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 1–9. York: White Rose University Press, DOI: https://doi.org/10.22599/White.a.

Physicalism is the modern anti-Cartesian theory of what the mind is. This view says that everything that exists is either itself a physical entity or somehow constituted by, composed by, or exhaustively determined by physical entities. The main draw of physicalism about the mind is that it seems, at first, to easily solve the problem of mental causation. Physicalism says that when we talk about someone's mental life we are actually talking about physical states, properties or events, so mental causation reduces to causation by certain physical states, properties or events. In its crudest form, this kind of physicalism says that mental states and events are neural states and events, and mental causation is causation by neural states and events.

As it turns out, the most popular kind of physicalism has difficulty dealing with the problem of mental causation. Most contemporary philosophers of mind who call themselves physicalists accept some form of 'non-reductive physicalism'. On this view, that which is mental is not *identical* with anything physical; nevertheless, physical states, events and properties realise, constitute, compose or exhaustively determine mental states, properties and events. This kind of physicalism is thought to be difficult to reconcile with the principle of the causal closure of the physical world, which says that 'at every time at which a physical event has a cause it has a sufficient physical cause' (Gibb 2013: 2). As Jaegwon Kim (2005) argues, if some physical events have mental causes and those mental causes are not identical with any physical entities (as non-reductive physicalism maintains), then these physical effects must be overdetermined or the principle of causal closure must be false.[2] Since this objection was raised, non-reductive physicalists have offered many counterarguments aiming to show that their version of physicalism can save the phenomenon of mental causation while respecting causal closure. For example, Karen Bennett (2003), Sydney Shoemaker (2013) and Steinvör Thöll Árnadóttir and Tim Crane (2013) argue that both mental entities and the physical entities that realise them can be causally efficacious without this being a case of 'double-causing' anything like the paradigmatic cases of overdetermination. The debate about whether non-reductive physicalism can solve the problem of mental causation or if a fully reductive version of physicalism is required is ongoing.

The aforementioned debate notwithstanding, physicalism remains a popular metaphysics of mind because it appears to be the only metaphysics of mind that can (a) permit mental causation and (b) respect plausible principles about what actual causation is like, such as the principle of causal closure. This argumentative strategy underlies the main argument for physicalism about the mind, which is known as the causal argument for physicalism. Debates within philosophy of mind tend to centre on which kind of physicalism gives the best reconciliation between (a) and (b). Non-physicalist alternatives are generally thought incapable of giving any kind of reconciliation at all. In this way, contemporary philosophy of mind is shaped by this question: how is it possible for

[2] See also Crane (1995) and Heil (2013).

the mental to causally interact with the physical, especially given the apparent physicality of causation?

However, I believe that this is the wrong question to ask. I believe that contemporary philosophy of mind labours under a misapprehension of what mental causation is.

In most discussions of the problem of mental causation, mental causation is presented as a cause–effect relation between mental and physical entities. In many cases, mental causation is presented as a causal relation between mental and physical *events*. Sometimes mental causation is presented as a causal relation that can hold between *states*. Less frequently, mental *processes* are mentioned. Usually, events, states and processes are thought of as being very similar in nature, so that there is no need to treat mental events, mental states and mental processes differently when considering their candidacy as causal relata. In most discussions of mental causation, events, states and processes are thought of as three subclasses of the same general ontological category. Members of this general ontological category—I will call them *items*—are typically thought of as *particulars*, where particulars are unrepeatable, concrete individuals. So, even where mental causation is not presented as a causal relation between events—or not only between events—it is still presented as a causal relation between items that are mental and items that are physical. I call this understanding of mental causation the relational understanding of mental causation:

> **Relational understanding of mental causation:** mental causation is mental items (events, processes or states) standing in causal relations to physical items (e.g. movements of a person's body).

Central to the relational understanding of mental causation is the idea that mental causation is a cause–effect relation between mental and physical items; mental phenomena are thought of as links in causal chains. This is the understanding of mental causation that has become standard in philosophy of mind but which I think is misconceived.

I believe that the relational understanding of mental causation is presupposed in many debates within philosophy of mind because of a triad of philosophical theories: (1) physicalism, (2) causal theories of intentional action and (3) relational approaches to causation. Although these theories are logically independent and about distinct philosophical questions, in practice they are mutually reinforcing. The relational understanding of mental causation presupposed by most arguments for physicalism is made to seem indispensable because of causal theories of intentional action, which in turn owe much of their apparent plausibility to relational assumptions about causation, assumptions that physicalists are likely to make. I believe this triad of views has limited our thinking about mental causation and therefore prevented us from exploring more diverse accounts of the relationship between our mind and body. I call this triad of views *the physicalist triad* because the upshot of endorsing each

element of the triad is that physicalism becomes the only acceptable metaphysics of mind as it appears to be the only view that has a chance of saving the phenomenon of mental causation. The aim of this book is to try to break out of this triad in order to open up new ways of understanding mental causation and thereby refresh debates within philosophy of mind.

I am not the first to suggest that there are connections between philosophy of action, philosophy of causation and physicalism. E. J. Lowe (2008) argues that physicalist consensus in philosophy of mind prevents and undermines a powerful account of rational agency. Jennifer Hornsby (2015) also argues that neo-Aristotelian theories of action—the main rivals to causal theories of action—call into question the existence of the kind of mental causation that forms the subject of debate in philosophy of mind, and hence have consequences for causal arguments for physicalism. However, Hornsby points out that 'none of this work appears to have made any impression upon work in mainstream philosophy of mind' (2015: 133). I suspect this is because the connection between physicalism, causal theories of intentional action and relational approaches to causation has not been sufficiently explicit to those working within philosophy of mind. Furthermore, no-one has provided reasons to persuade someone dissatisfied with the causal argument for physicalism that their best strategy for resisting the conclusion of this argument is to use lessons from philosophy of action and causation to question the foundational assumption of the causal argument. This is what I intend to do.

The arguments I put forward here will be of interest to those who are sceptical of physicalism as a metaphysics of mind but also feel dissatisfied with the standard counterarguments to physicalism. What I offer here is a distinctive non-physicalist approach to the problem of mental causation. However, I will not argue directly against physicalism. Ultimately, it is the relational approach to causation, and not physicalism itself, that does the most harm to our understanding of mental causation. Nevertheless, I hope to provide reasons to question physicalism's hegemony as the metaphysics of mind that best accommodates mental causation.

In my view, the dominance of physicalism in philosophy of mind is not indicative of physicalism's veracity. Instead, it ought to be something to make us suspicious. Physicalism is commonly thought of as the only naturalistic metaphysics of mind. Alternatives to physicalism are quickly criticised for rendering our mental lives inefficacious or for being at odds with scientific understanding. Physicalism has also become the theoretical backdrop for many of the kinds of questions discussed within contemporary philosophy of mind, such as: how do thoughts cause behaviours? what are the neural correlates of consciousness? how are mental entities and physical entities related if they are not identical? In this way, physicalism has prescribed what kinds of questions we ask about action, mental causation and the mind–body connection. This suggests to me that we need to interrogate the ideas about mental causation that contemporary philosophy of mind is taking for granted and which make physicalism seem like the only option.

In this book, I argue that physicalism's dominance, and the dismissal of non-physicalist alternatives as unnaturalistic or unscientific, depends on an understanding of mental causation that is not as theory–neutral as it first appears and relies upon (as it turns out) questionable assumptions about causation. My aim with this book is to provide a different, hopefully more philosophically neutral, description of the mental causation associated with intentional action. In this way, I hope to give us a fresh starting point for developing an alternative metaphysics of mind and for asking new questions about action, mental causation and the mind–body connection.

This book is divided into two parts. The first part explores the views that make up the physicalist triad. I explain how the three views are interconnected and provide evidence that, while logically independent, the views are mutually reinforcing. I also explain how these three views are responsible for the widespread acceptance of the relational understanding of mental causation. The philosophical topics discussed in these chapters will probably be familiar to the reader. However, it is my hope that by examining the interconnections between physicalism, causal theories of intentional action and relational approaches to causation I can reveal some important, but often unstated, assumptions made by these theories.

In Chapter 1, I outline physicalism in more detail and explain how arguments for physicalism presuppose the relational understanding of mental causation. I also explain how physicalism is supported by the other two elements of the physicalist triad.

In Chapter 2, I outline causal theories of intentional action. These theories have their roots in work by Donald Davidson. Davidson (1963) argues that when we say someone acted as they did because they wanted to do something, or because they believed that something was the case, we are giving a causal explanation. From this, Davidson concludes that states of desiring and states of believing—or at least events suitably related to states of desiring and believing, such as the onset of the desire or belief—are causes of the actions they explain. This argument has inspired the view that intentional actions are events that are caused by mental items. I explain how this view is used to justify the relational understanding of mental causation. I also argue that causal theories of intentional action owe much of their plausibility to relational approaches to causation.

In Chapter 3, I explain what a relational approach to causation is. A theory of causation is relational if and only if it is committed to the following thesis:

> **Relationalism:** causation is always and everywhere a relation; the worldly phenomenon that is referred to by our concept 'causation' is not ontologically diverse in this respect.

The regularity theory of causation and David Lewis's (1973a; 1973b) counterfactual theory of causation are paradigm examples of relational theories of causation. However, there are many other examples.

In Chapters 4 and 5 I explain why I think we ought to challenge the physicalist triad. I do not argue directly against any of the theories that make up the triad. I do not argue that physicalism fails on its own terms, or that the causal theory of action cannot tell us what intentional action is, or that a relational theory of causation is impossible. Instead, I focus on what I take to be the weakest point of the triad, which is the account of agency it provides. In Chapter 4, drawing on arguments presented in philosophy of action, I argue against the physicalist/event-causalist description of agency provided by the physicalist triad. In Chapter 5, I offer a critical examination of some existing alternative theories of agency that appeal to the concept of agent causation or substance causation. I suggest that the chief failing of these theories is that they do not go far enough when it comes to rejecting the relational approach to causation.

In the second part of this book I show how broadening our understanding of causation, and more specifically incorporating the concept of *process* into our understanding of causation, opens up new ways of understanding intentional action and the mental causation associated with it. In this way, I hope to describe what a theory of mental causation can look like if the physicalist triad is rejected. I provide reasons to think that this alternative approach to causation allows us to develop a better understanding of intentional action and the mental causation associated with it.

In Chapter 6, I present my own non-relational approach to causation. My approach denies that causation is always a relation and holds instead that causation can be a process rather than a relation, of which processes like breaking, crushing, bending etc. are more determinate species. My proposal is that causation is on display not only when events make the difference to the occurrence of other events but also when substances engage in processes. I suggest that engaging in a process is analogous to instantiating a property, and that events are instances of processes.

In Chapters 7 and 8, I challenge Davidson's argument that states of desiring and states of believing are causes of the actions they explain. This argument has been challenged before. Philosophers such as Ludwig Wittgenstein (1958) and Elizabeth Anscombe (1957) rejected the idea that beliefs and desires stand to actions as causes to effects. They argued that concepts like *belief, desire* and *intention* do not refer to items that can stand in causal relations to actions or physical events. Similarly, Gilbert Ryle (1949) argued that 'mental conduct verbs'—like 'knowing', 'believing', 'intending' and 'desiring'—do not signify or denote inner causal events, so when such verbs are employed to explain why an agent acted they do not designate inner causes of the action they explain. This view, which I call the *non-causalist* view, denies that intentional action entails the existence of causal relations between mental items and physical events.

However, non-causalists reach this conclusion by arguing that explanations of intentional actions that cite beliefs or desires are not usually causal explanations at all, whereas I believe that explanations of intentional actions that cite the agent's beliefs or desires do give causal information. Fortunately, this kind

of intermediary view is made possible if one rejects the relational approach to causation. In Chapter 7, I argue that it is not necessary for an explanation to be causal that its explanandum designate an effect and its explanans designate an item that is the cause of that effect. My non-relational theory of causation implies that facts about causal relations are not the only causal facts that causal explanations could answer to. I suggest that some causal explanations are made true by the non-relational aspect of causal reality, that is, by facts about substances engaging in processes.

In Chapter 8, I argue that explanations of intentional action that cite the agent's reasons for acting are the kind of causal explanation that are not made true by causally related events. The most important consideration favouring this view is that it saves two strong intuitions: (a) that reason-giving explanations are causal and (b) that the mental states cited in reason-giving explanations do not denote items that stand in causal relations to the actions they explain. This view has important consequences for how we ought to think about the nature of intentional action. Most importantly, it casts doubt on the view that intentional actions are distinguished from non-intentional actions by their causes.

In Chapter 9, I propose an alternative view of intentional action. I propose that to act intentionally is to engage in a process, and as such is to exercise a power—but a power of a special sort. The power to act intentionally is a power to structure one's own activities so that they demonstrate a pattern—a pattern that is only revealed by attributing mental states to the agent. So, when an agent acts intentionally, they engage in the process of causation. The process they engage in counts as *mental* causation in virtue of the fact that the agent is manifesting a special power to organise their activities so that they instantiate a certain structure, a structure that is made comprehensible by the agent's mental states.

In Chapter 10, I revisit the problem of mental causation. If the arguments of the previous chapters are successful, then the existence of intentional action does not entail that mental items stand in causal relations to physical items. When we say that someone acted intentionally because of what she believed, desired, intended or decided, these mental concepts need not refer to items that stand in causal relations to physical events. Instead, it is possible to think of the mentality of the causal processes human beings engage in when they act intentionally to consist in the fact that these processes are part of a larger pattern of meaningful, or interpretable, activity. This means that the standard way mental causation is set up as a problematic subject in philosophy of mind may not be right. As explained above, debates within philosophy of mind tend to centre on which metaphysics of mind best reconciles the claim that mental items stand in causal relations to physical items with plausible principles about what actual causation is like, such as the principle of causal closure. However, if realism about mental causation does not require the relational understanding of mental causation at all, then the problem of mental causation as it is usually understood is a pseudo-problem. In Chapter 10, I discuss alternative ways

to understand mental causation and the consequences this has for philosophy of mind.

I think it is undeniable that our mental life makes a difference to our bodily life. I agree that what we do with our bodies causally depends on what's going on in our minds. However, I think it has been a mistake to assume that the causal aspect of the mind–body connection ought to be understood as causal interaction between mind and body. Descartes was wrong, I believe, to divide human beings into two distinct substances, mind and body. Modern philosophers of mind are similarly wrong to divide mental causation into a causal exchange between distinct aspects of ourselves.

References[3]

Anscombe, G E M 1957 *Intention*. Oxford: Basil Blackwell.

Árnadóttir, S T and Crane, T 2013 There is no exclusion problem. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 248–266.

Bennett, K 2003 Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs*, 37(3): 471–497. DOI: https://doi.org/10.1111/1468-0068.00447

Crane, T 1995 The mental causation debate. *Aristotelian Society Supplementary Volume*, 69(1): 211–236. DOI: https://doi.org/10.1093/aristoteliansupp/69.1.211

Davidson, D 1963 Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700. DOI: https://doi.org/10.2307/2023177. Reprinted in Davidson 2001 pp. 3–20.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Gibb, S 2013 Introduction to mental causation and ontology. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 1–17.

Heil, J 2013 Mental causation. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 18–35.

Hornsby, J 2015 Causality and 'the mental'. *HUMANA.MENTE Journal of Philosophical Studies*, 8(29): 125–140.

Kim, J 2005 *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.

Lewis, D K 1973a Causation. *The Journal of Philosophy*, 70(17): 556–567. DOI: https://doi.org/10.2307/2025310

[3] Author note: some references to Davidson are formatted (1963/2001). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001 collection of his essays, with the page numbers relating to that volume.

Lewis, D K 1973b *Counterfactuals*. Oxford: Basil Blackwell.
Lowe, E J 2008 *Personal agency: The metaphysics of mind and action*. New York: Oxford University Press.
Ryle, G 1949 *The concept of mind*. London: Hutchinson's University Library.
Shapiro, L (ed.) 2007 *The correspondence between Princess Elisabeth of Bohemia and René Descartes*. Chicago, IL: University of Chicago Press.
Shoemaker, S 2013 Physical realization without preemption. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 35–57.
Wittgenstein, L 1958 *The blue and brown books*. Oxford: Blackwell.

PART I

The Physicalist Triad

CHAPTER 1

Physicalism

In this chapter I introduce the first element of the physicalist triad. Physicalism is the view that everything that exists is either itself physical or something composed of, constituted by, realised by, grounded in, or in some other way 'nothing over and above' physical entities. Physicalism is a popular account of the mind. Very roughly, physicalism about the mind says that mental entities (states, properties, events, processes) are nothing over and above physical entities, which are brain states or brain activity on some versions of physicalism. The strongest argument for physicalism takes the existence of mental causation as a premise, and concludes that mental causation can only be possible, given certain assumptions about what causation in the actual world must be like, if some form of physicalism is true. Versions of this argument include the 'causal argument for physicalism' championed by David Papineau (1993; 2001; 2002), Donald Davidson's (1970) argument for anomalous monism, and Jaegwon Kim's (1993; 1998; 2001) 'causal exclusion argument'.

In what follows, I outline these arguments in detail. My aim in this chapter is not to challenge these arguments by disputing their premises. Instead, my aim is to reveal some important assumptions that are implicit in these arguments. The most important of these is an assumption of what mental causation is. Arguments for physicalism assume that mental causation is mental items (events, processes or states) standing in causal relations to physical items (e.g. movements of a person's body). I call this the relational understanding of mental causation. I aim to show in Section 1.2 that this assumption is essential to arguments for physicalism: without it the arguments are invalid. I also argue, in Sections 1.3 and 1.4, that the plausibility of the relational understanding of mental causation in turn derives from two other philosophical theories: causal theories of intentional action and relational approaches to causation. According to causal theories of intentional action, intentional or voluntary human action is possible only if mental items stand in causal relations to physical events such as bodily movements. According to relational approaches to

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 13–37. York: White Rose University Press. DOI: https://doi.org/10.22599/White.b.

causation, any naturalistic account of causation is committed to the idea that all causation is the same; causation is a homogenous phenomenon. Causation can hold between diverse relata, but there is one relation of causation, and that is all causation is. On this approach, what makes an example of causation 'mental' can only be that one or both of the causally related entities is mental.

In the last section of this chapter, Section 1.5, I outline reasons why someone might be dissatisfied with physicalism as a metaphysics of mind. I do not intend to disprove physicalism in this section. I only intend to show that physicalism is not without its critics. The causal arguments for physicalism that I describe in Section 1.1 can make it seem like physicalism is the only option. Showing that there are some reasons to doubt the truth of physicalism helps justify my project, which is to show that the causal arguments for physicalism are artificially bolstered by their association with causal theories of intentional action and relational approaches to causation.

1.1 Physicalism and mental causation

In philosophy of mind, the problem of mental causation is the problem of how that which is mental is able to causally interact with that which is physical. This problem is usually presented as a 'how possibly' question, that is, a question about how mental causation could exist. It is usually accepted as prima facie true that mental causation does exist; the difficulties arise only when we try to understand how this could be, given certain assumptions about the fundamental nature of reality. As Peter Menzies states, 'philosophical questions about mental causation revolve around … how it is possible in the first place in the light of certain metaphysical assumptions and principles' (2013: 58). The metaphysical assumptions and principles that seem to make the existence of mental causation puzzling concern the apparent physicality of the causal world. For example, Kim tells us that the problem of mental causation is 'to explain how mentality can have a causal role in a world that is fundamentally physical' (2005: 1). The existence of mental causation is thought to be especially difficult to reconcile with the principle of the causal closure of the physical world, which says that 'at every time at which a physical effect has a cause it has a sufficient physical cause' (Gibb 2013: 2).[4]

[4] There are many alternative formulations of the principle of causal closure, for example: 'any physical state or change, if it has a cause or explanation, has a physical cause or explanation' (Hopkins 1978: 223); '[f]or all physical events and states there are necessary and sufficient physical conditions, their "explanations" or "causes"' (Skillen 1984: 514); '[e]very physical effect has its chance fully determined by physical events alone' (Noordhof 1999: 367); and '[a]ll physical effects are fully determined by law by prior physical occurrences' (Papineau 2001: 9). These various formulations are

Proposing a metaphysics of mind that reconciles the existence of mental causation with principles about what causation in the actual world must be like, such as the principle of causal closure, has been the aim of a substantial amount of research in philosophy of mind. A key aim in contemporary philosophy of mind is to establish a metaphysics of mind that (a) saves the phenomenon of mental causation and (b) respects plausible principles about what actual causation is like, such as the principle of causal closure. Physicalism, of some form or another, is generally considered to be the metaphysics that has the best chance of satisfying these two conditions.

Physicalism says that everything that exists is either itself a physical entity or 'nothing over and above' a physical entity. Exactly what it is for one entity to be 'nothing over and above' another is a matter of debate. Some physicalists express their position in terms of *supervenience* (Haugeland 1982; Hellman and Thompson 1975). This version of physicalism says that all the non-physical aspects of the world supervene on the physical aspects, which is to say that it is impossible for two things to differ with respect to their non-physical properties without also differing in their physical properties. There are a number of difficulties associated with expressing physicalism in terms of supervenience.

First, supervenience can be global or local. Physicalism expressed in terms of global supervenience says that no two *worlds* can differ with respect to their non-physical properties without also differing in their physical properties. As Frank Jackson puts it, '[a]ny world which is a minimal physical duplicate of our world is a duplicate of our world' (1998: 14). Physicalism expressed in terms of local supervenience says that no *person/event/object* can differ with respect to their non-physical properties without also differing in their physical properties. For example, Davidson claims that supervenience of the mental on the physical amounts to the claim that 'there cannot be two events alike in all physical respects but different in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect' (1970/2001: 214).

One reason to opt for global supervenience rather than local supervenience is that global supervenience accommodates externalism about the content of mental states. Externalism holds that the content of mental states is determined in part by the nature of the environment and not by the intrinsic properties of the thinker. For example, suppose Oscar thinks 'water is wet'. Externalism says that, for Oscar's thought to be the thought that it is, Oscar must be related to

---

not equivalent. Furthermore, the principle of causal closure is supposed to be derived from, and supported by, the findings of scientific investigations into causal processes, but it is a contentious question which, if any, of the various formulations of causal closure enjoy such support. See Lowe (2000) for a discussion of issues relating to the principle of causal closure, and see Papineau (2001) for a defence of the principle.

certain external states of affairs. For Oscar's thought to be about Earth-water, the stuff occupying rivers that is composed of hydrogen and oxygen, Oscar's environment must contain Earth-water or his linguistic community must have been in contact with Earth-water. Twin-Oscar, who lives in an alternative world that does not contain Earth-water but instead contains a substance with exactly the same observable properties but which has a different molecular structure, cannot have the same thoughts as Oscar. This is so even if both Oscar and Twin-Oscar know nothing about the molecular structure of the watery stuff that exists in their worlds and are disposed to ascribe the exact same properties to the watery stuff of their worlds. Externalism is a plausible and well supported thesis (Burge 1979; Putnam 1975) so it would be good for physicalism to be consistent with it. However, the difficulty with expressing physicalism in terms of global supervenience is that the theory does not seem to be meaningfully different to non-physicalist views of the mind such as emergentism, the view that the mental aspects of the world depend on the physical aspects of the world but have their own independent causal powers (thus causal closure is false) (Horgan 1993; Wilson 2005). Expressing physicalism in terms of local supervenience makes physicalism a more substantive thesis, but one that has a higher chance of being false. Externalism about mental contents seems to disprove it: even if we assume that Oscar and Twin-Oscar have all the same intrinsic physical properties, the fact that they live in different worlds entails that they have different thoughts.

Another issue with expressing physicalism in terms of supervenience is that supervenience captures covariance between supervenient and subvening properties but that is all. Therefore, to claim that everything that exists supervenes on the physical is not very informative. To say that mental properties supervene on physical properties does not tell us very much about how the mental and the physical relate. If one property can be shown to be identical to another, then there will be a relation of supervenience between them. However, supervenience also holds between properties related by realisation, constitution, or a determinable–determinate relation. What these relations have in common is that they are all one-way necessitation relations: the subvening property necessitates the presence of the supervenient property but not the other way around. As Terence Horgan argues, if physicalists express their view in terms of supervenience they must also explain why supervenience holds between the physical and the non-physical in a 'materialistically acceptable way' (1993: 556). In other words, physicalists need to elucidate the relation between the physical things and the putative non-physical things that explains why supervenience holds, and why the physical things are more fundamental or more real or, in some other way, ontologically superior. As Amanda Bryant puts it, 'supervenience formulations of physicalism fail to capture or illuminate the common physicalist contention that the mental metaphysically depends on the physical' (2020: 489). In response to this challenge, some physicalists have expressed their view in terms of realisation: non-physical properties are

realised by physical properties (Melnyk 2003; Melnyk 2006; Melnyk 2018; Shoemaker 2001; Shoemaker 2007); others in terms of constitution (or an analogue of constitution): non-physical properties are constituted by physical properties (Pettit 1993); others in terms of a determinable–determinate relation: non-physical properties are determinable properties of which physical properties are the determinates (Yablo 1992); and others in terms of fixing: non-physical properties fix all other properties (Elpidorou & Dove 2018); truthmaking: physical facts make true all facts (Morris 2018); and, more recently, grounding: non-physical properties and facts are grounded by physical properties and facts (Bryant 2020). It would take us too far afield to assess each of these proposals individually. The debate concerning what it means to say that everything that exists is nothing over and above the physical is ongoing, with some writers more pessimistic than others about whether the physicalist can adequately meet this challenge (examples of pessimists include Horgan (1993), Lynch and Glasgow (2003) and Wilson (2016)).

Regardless of how the nothing-over-and-above relation should be spelt out, several arguments conclude that some form of physicalism about mentality must be true, because otherwise an account of mental causation that respects plausible principles about what causation is like (such as causal closure) is impossible. Three such arguments are the 'causal argument for physicalism' championed by Papineau (1993; 2001; 2002),[5] Davidson's (1970) argument for anomalous monism, and Kim's (1993; 1998; 2001) 'causal exclusion argument'.

As Papineau (2001: 9) presents it, the causal argument for physicalism has three premises:

1. 'All mental occurrences have physical effects'.
2. 'All physical effects are fully determined by law by prior physical occurrences' (this is Papineau's formulation of the causal closure principle).
3. 'The physical effects of mental causes are not all overdetermined'.

From these three premises, it is concluded that '[m]ental occurrences must be identical with physical occurrences' (2001: 9). The causal argument is used to establish a physicalist identity theory about mental *occurrences* or *events*. However, proponents of the causal argument typically assume that events, processes and states are not significantly different in nature, and so what goes for mental *events* goes for mental *processes* and mental *states* (or at least 'token states') as well. Importantly, the first premise is understood by proponents of the causal argument as simply equivalent to the claim that mental causation exists.

Davidson's argument for anomalous monism is another example of an argument for identifying mental events with physical events that takes the

[5] This argument is also known as 'the causal overdetermination argument' (Crane 1995; Gibb 2013) and 'the overdetermination argument' (Noordhof 1999; Sturgeon 1998).

existence of mental causation as a premise. Anomalous monism asserts that every individual mental event is identical with some physical event, but mental kinds are distinct from physical kinds. So, for example, even though *being a decision* cannot be identified with any physical kind, not even *being a brain event*, every token decision is identical with a physical event of some kind. Davidson's (1970/2001: 208) argument for this view involves three premises:

1. 'At least some mental events interact causally with physical events' (the 'principle of causal interaction').
2. 'Where there is causality, there must be a law: events related as cause and effect fall under strict deterministic laws' (the 'principle of the nomological character of causality').
3. 'There are no strict deterministic laws on the basis of which mental events can be predicted and explained' (the 'anomalism of the mental').

Again, the first premise of this argument is typically understood as an assertion that mental causation exists.

Kim objects to Davidson's anomalous monism on the grounds that 'on anomalous monism, events are causes or effects only as they instantiate physical laws, and this means that an event's mental properties make no causal difference' (1989: 34–35). On Kim's view, events are causes in virtue of the properties they involve, and not every property an event involves causally matters, unless there is causal overdetermination. Kim argues that the nomological character of causality and anomalism of the mental imply that it is always an event's *physical* nature that is causally relevant; whatever *mental* properties an event may involve are excluded from being causally relevant by physical properties, which enjoy superior candidacy for this status. This is Kim's 'causal exclusion argument', which is another example of an argument for a particular metaphysics of mind—this time a physicalist identity theory that identifies mental *properties* with physical *properties*—which focuses on the problem of mental causation.

Kim's causal exclusion argument has also been directed against 'non-reductive' physicalist views. According to non-reductive physicalism, that which is mental is not *identical* with anything physical; nevertheless, physical states, events and properties realise, constitute, compose or exhaustively determine mental states, properties and events. In other words, a one-way necessitation relation obtains between physical and mental states, events and properties. Kim (2005) argues that, if mental entities are both distinct from physical entities and causally responsible for some physical effects, then these physical effects must be overdetermined or the principle of causal closure must be false.[6] Non-reductive physicalists have responded by questioning Kim's key assumption that an effect cannot have two sufficient causes unless there is overdetermination. According

[6] See also Crane (1995) and Heil (2013).

to these responses, because of the one-way necessitation relation between mental and physical entities, it is wrong to see them as overdetermining causes of the same effect (Árnadóttir and Crane 2013; Bennett 2003; List and Menzies 2009; Shoemaker 2013). These responses grant that, if physical states, events and properties did not realise, constitute, compose or exhaustively determine mental states, properties and events then there would be overdetermination. Therefore, these responses assume that some form of physicalism is still required to reconcile the phenomenon of mental causation with principles about what actual causation is like, in this case the principle of causal closure and the assumption that causal overdetermination is not prolific. In this way, considerations about mental causation are still taken to support some kind of physicalism.

1.2 The relational understanding of mental causation

The arguments for physicalism outlined above are strong. It seems like the only way to deny their conclusion is to reject causal closure, accept prolific overdetermination or deny that there is mental causation. Few are willing to reject causal closure as this principle is thought to enjoy empirical support and prolific overdetermination is widely regarded as implausible. This seems to leave epiphenomenalism—the view that mental events are caused by physical events, for example events occurring in the brain, but mental events have no physical effects—as the only option left for the non-physicalist. However, before we declare physicalism the only acceptable metaphysics of mind, I think it is important to question whether the conception of mental causation assumed by arguments for physicalism is the right way to think about the place of mentality in the causal world.

In most discussions of the problem of mental causation, mental causation is presented as a causal relation between mental and physical *events*. Recall Davidson's principle of causal interaction: 'at least some mental events interact causally with physical events' (1970/2001: 208), or the following formulations of the first premise of the causal argument: 'all mental occurrences have physical effects' (Papineau 2001: 9); 'we think of mental and physical events as causally related' (Hopkins 1978: 223); and '[m]ental events have physical effects' (Noordhof 1999: 367). Sometimes, mental causation is presented as a causal relation that can hold between *states*. For example, in Anthony Skillen's (1984: 514) version of the causal argument for physicalism, he claims that '[o]f some physical events and states, mental events and states are causes'. Often, events, states and processes are thought of as being very similar in nature. For example, when David Armstrong proposes that mental states are states that are 'apt for bringing about a certain sort of behaviour', he notes that his use of the word 'state' is 'not meant to rule out "process" or "event"' (1968: 82). In most discussions of mental causation, events, states and processes are

thought of as three subclasses of the same general ontological category, and any differences between them do not affect their suitability to be the relata of a causal relation. I will call members of this general ontological category *items*. In most discussions of the problem of mental causation, mental causation is presented as a cause–effect relation where at least one of the relata is a mental item. I call this understanding of mental causation the *relational understanding of mental causation*.

> **Relational understanding of mental causation:** mental causation is mental items (events, processes or states) standing in causal relations to physical items (e.g. movements of a person's body).

What is important to notice about the relational understanding is that, as Jennifer Hornsby puts it, it presents mental causation as something 'we are supposed to think of as causation *by* the mental' (2015: 129). Or, as Tim Crane puts it, 'the arguments for physicalism must assume that the labels "mental" and "physical" as applied to causation are really transferred epithets—what is mental and physical are the relata of causation, not the causation itself' (1995: 219). Most discussions of mental causation thus assume that what is distinctive about mental causation is that it involves a mental relatum. For example, Thomas Kroedel writes that mental causation is 'the causation of physical effects by mental causes … there is mental causation whenever what is going on in our minds causes our bodies to move' (2020: 1). Philosophers writing about the problem of mental causation are limited to this way of describing what mental causation is because they assume that 'cause' is an unequivocal term—all causation everywhere is the same kind of thing, so the only thing that can discriminate between different categories of causation is the nature of the relata involved. This is the understanding of mental causation that has become standard in philosophy of mind. However, I believe this understanding of mental causation is misconceived.

The relational understanding of mental causation encourages us to accept an ontology of mental *items*. It presupposes that mental concepts pick out items, which can be causes and whose intrinsic nature is up for discovery. So, when someone believes something, an item called a belief exists. When someone wants something, an item called a desire exists. However, it is not obvious to me that the best way to describe mentality is as a collection or succession of mental items. I am sceptical that our status as minded creatures depends on the existence of mental items whose nature we have yet to discover and whose existence must, one way or another, be reconciled with the idea that the world is physical in all its fundamental aspects. On this matter, I agree with Gilbert Ryle. Ryle (1949) objects to what he calls the 'para-mechanical' view of the mind. According to this view, to have a mind—to have beliefs, desires, intentions, ideas, ambitions, emotions etc.—is to possess an inner causal mechanism that operates invisibly but which has observable actions as output. Mental states and

processes are conceived of as hidden inner causes of observable behaviour, and this, for Ryle, is a mistake. Ryle argues that when we use mental predicates we are not denoting (or even connoting) episodes in a person's secret history or alterations to an inner stream of consciousness. Furthermore, when we say that someone acted because of what she believed, knew, intended, desired, imagined, remembered or felt, these concepts do not—in fact, *could* not—refer to items that stand in causal relations to physical events. The idea that there can be mental causes is a category mistake, according to Ryle. Mental states and processes are just not the right sort of thing to be causes and effects.

I agree that mental states and processes are not the right sort of thing to be causes and effects. At the very least, it is wrong to assume that they can be causes in anything like the same sense in which the strike of a match can be the cause of its lighting, or the dropping of a glass can be the cause of its breaking. One reason for this is because it is a mistake to gloss over the differences between events, states and processes and assume that all three kinds of entity can be causes in the same way.

One assumption that is often made is that states and processes—or at least token states and token processes—are *particulars* just like events. However, this assumption is questionable. Particulars are *concrete, unrepeatable* entities. Helen Steward (1997) further suggests that something is a particular if it is capable of having a 'secret life' where an entity has a secret life if and only if:

1. the entity might be uniquely identified by means of some referring expression which is not known to apply to it by someone who is, nevertheless, in a position to single that entity out in some other way;
2. for some such referring expressions, the subject's not knowing that they provide an alternative means of uniquely identifying the entity in question is not simply a matter of her being ignorant of an alternative means of uniquely identifying some other entity;
3. for some such referring expression, the subject not knowing that they provide an alternative means of uniquely identifying the entity in question is not simply a matter of her not knowing about one of the entity's relational properties (where spatial and temporal properties are not accounted relational). (1997: 32)

Entities that satisfy these conditions are entities that can be uniquely referred to with more than one expression, each of which picks out the entity via a different intrinsic feature of it. For example, the morning star and the evening star are one and the same entity, the planet Venus, but each expression picks out Venus via different intrinsic features of it: appearing in the sky in the morning and appearing in the sky in the evening. It is because each expression refers to Venus via different intrinsic features that identifying the morning star with the evening star is informative and not trivially true. Venus satisfies the secret life requirement; therefore, it is a particular.

Steward argues that entities that have a propositional structure, like facts, cannot be particulars because they do not meet the secret life requirement. Steward (1997) argues convincingly that states also have a propositional structure and so are also not particulars. I also do not think that processes are particulars. I think they are universals—they are single repeatable entities. I will have more to say about this in Chapter 6 (see also White 2020). This means that, even if mental states and processes are cited in causal explanations, we cannot assume that the role they play in such explanations is exactly the same as the role we would take an event to play.

Denying the existence of any mental particulars at all is too strong to be plausible. Suddenly remembering something, successfully imagining something, noticing something, realising something all seem to be mental events and all seem to exist. However, in Rylean spirit, I think these events should not be thought of as occurring in a 'secret history'. Such events are often thought of as episodes of a mental succession—and this is what I believe is wrong. Although these events are mental, they are, in a very banal sense, physical events too because they are all actions of human beings. These events should be dealt with as actions—not as happenings in a private mental sequence, whose intrinsic nature is unknown to us.

I do not want to say there is no such thing as mental causation. I am sceptical of the idea that there are mental *items* that stand in causal relations but I do think there is causation that deserves to be called 'mental'. I believe that the causal processes human beings engage in when they act intentionally count as mental causation, and the mentality of these causal processes consists in the fact that these processes are part of a larger pattern of meaningful, or interpretable, activity. One aim of the later chapters of this book is to prove this.

An example might help to summarise why I am sceptical of the relational understanding of mental causation. Some years ago, my partner was driving along a country road at night when, from the passenger's seat, I spotted a deer in the road in front of us. I immediately called out "deer, deer, deer!" to alert the driver. I have three intuitions about this example. First, the idea that the event-causal sequence from first noticing the deer to calling out involves a mental-but-not-physical event as a causal intermediary seems incorrect. My intuition is that, if we were to describe the example in event-causal terms we would mention light entering my eye, electrical activity in my brain, contractions of my muscles—all events that we would class as physical—and that's all. Second, the idea that one of the physical events in the chain just described, or even a subset of those physical events, is identical with, or somehow constitutes the mentality of the example also seems wrong. It seems wrong to suppose that any of the physical events in the causal sequence amounts to my conscious experience, or to my worry that we might hit a deer, or to my desire that the driver be alerted, or any of the mental experiences I think are present in the example. (In other words, I do not think physicalism is true.) Third, I think there *is* mental causation in this case—I did as I did because of what I thought and experienced.

The puzzle is: how can all these intuitions be true? The solution, I suggest, is to rethink mental causation. What makes this case an example of mental causation is not that there is a causally efficacious mental item that caused my behaviour.

1.3 Mental causation and human agency

I believe that the relational understanding of mental causation is presupposed in many debates within philosophy of mind because of a triad of popular, and individually plausible, philosophical theories: physicalism, causal theories of intentional action and a relational approach to causation. I call this triad of views *the physicalist triad*. Although these theories are logically independent and about distinct philosophical questions, in practice they are mutually reinforcing.

To see this, we should first ask why the relational understanding of causation has become the standard way of thinking about mental causation within philosophy of mind. As I have already mentioned, few philosophers writing on the problem of mental causation are willing to deny the existence of mental causation. However, there is an important exception to this norm. 'Epiphenomenalism' is the view that mental items are caused by physical items, for example events occurring in the brain, but mental items have no physical effects. Epiphenomenalism can be seen as a response to the causal argument for physicalism. The argumentative force of the causal argument is that if mental and physical items are distinct, then they are in competition with each other for status as the cause of a physical effect, for example a bodily movement. The first premise of the causal argument states that mental items have such physical effects; the second premise says that every physical effect has a sufficient physical cause, so either the physical effects like bodily movements are overdetermined, or the mental and physical causes are one and the same, or we will have to accept that one of the candidate causes is not really a cause after all. It is often assumed that, if you are pushed towards the last option, you have to admit that it is the mental candidate that turns out not to be a cause. This is precisely what epiphenomenalists do.

Epiphenomenalism is a minority position, however. Most philosophers of mind take it as a position to be avoided. It will be useful to briefly consider why this is, as it provides evidence to think that the relational understanding of causation has become the dominant understanding of mental causation because of intuitions about human agency.

First, consider Kim's remarks on why it is important that mental causation is real:

> First and foremost, the possibility of human agency, and hence our moral practice, evidently requires that our mental states have causal effects in the physical world. In voluntary actions our beliefs and

> desires, or intentions and decisions, must somehow cause our limbs to move in appropriate ways, thereby causing the objects around us to be rearranged. (2005: 9)

Here Kim endorses the idea that the possibility of human agency depends on beliefs and desires, or intentions and decisions, standing in causal relations to bodily movements. I agree that some form of mental causation is indispensable to our conception of ourselves as agents who act intentionally and bear moral responsibility. I believe that such a conception presupposes the reality of 'causal processes involving mental phenomena', as Menzies (2013: 58) puts it. However, these claims are much weaker than the claim Kim makes. Kim claims that the possibility of human agency depends on beliefs, desires, intentions and decisions somehow causing our limbs to move.

Kim also states that:

> [I]t seems plain that the possibility of psychology as a science capable of generating law-based explanations of human behaviour depends on the reality of mental causation: mental phenomena must be capable of functioning as indispensable links in causal chains leading to physical behaviour, like movements of the limbs and vibrations of the vocal cord. A science that invokes mental phenomena in its explanations is presumptively committed to their causal efficacy; if a phenomenon is to have an explanatory role, its presence or absence must make a difference—a causal difference. (2005: 10)

Again, I agree that the worth of psychology as a science and as a means by which we can predict, explain and control each other's behaviour requires that people's behaviour can be causally explained by what they think, feel, believe and want. However, again, this is a weaker claim than Kim's. Kim claims that the possibility of psychological explanations of human behaviour requires that 'mental phenomena must be capable of functioning as indispensable links in causal chains leading to physical behaviour, like movements of the limbs and vibrations of the vocal cord'. Thus, Kim thinks the possibility of psychological explanation presupposes that mental phenomena, like believing that one ought to brush one's teeth or wanting to make a cup of tea, are *links in causal chains*.

Kim's justification for believing in mental causation seems to be that, unless mental items stand in causal relations to physical items, human agency and psychological explanation of human behaviour would be impossible. Kim also considers perceptual knowledge as evidence that mental causation—this time causation *of* mental events, as opposed to causation *by* mental events—must exist. The thought here is that if the content of perceptual experience is to indicate what the world is like then perceptual experiences must be caused by external states of affairs. However, this kind of mental causation is rarely the focus in debates about physicalism. Furthermore, this kind of mental

causation could not serve as a premise in the causal argument for physicalism described above. This is because the causal closure principle concerns the causes of physical *effects*, so the mental causation it potentially excludes is causation *of* physical effects *by* mental items.

It is not just Kim who thinks that human agency is possible only if mental items stand in causal relations to physical events such as bodily movements. The reaction to the work of Benjamin Libet shows that many others share Kim's belief. Libet (1985) conducted an experiment where participants were asked to move their finger when they felt like it and note the time at which they felt an urge to move their finger. While participants did this, Libet recorded their brain activity and found that a particular signal, known as a readiness potential, was correlated with the participants' finger movements and, significantly, occurred 350 milliseconds before participants reported having an urge to move their finger. This implied that the urge to move their finger could not have initiated the finger movement, and therefore that 'conscious will' does not play the role in 'voluntary action' that we think it does (Libet 1999; see also Libet 2002).

Libet thus assumes that for voluntary action to be possible, the action must be initiated by a mental item—in this case a conscious act of will. Similar work by John-Dylan Haynes and Michael Pauen (2013) found that neural activity correlated with the outcome of a choice whether to add or subtract digits occurred before participants recorded consciously making the decision. Haynes and Pauen take this to show that even more abstract choices, such as whether to do an addition or a subtraction, are not initiated by conscious mental events. Since Libet's early experiments, many more experiments on neurological preparatory processes for movement have been conducted, with neuroscientists asking what 'action-related cognitive processes' might be 'encoded' by such neurological activity (Fifel 2018: 785).

Libet's experiment, and those like it, have been criticised on the grounds that the task participants are being asked to do is artificial: participants are instructed to decide to do something spontaneously, hence extrapolation to other kinds of voluntary action, which might involve much more complicated or extended deliberation, is not justified. Other philosophers have argued that these kinds of experimental findings pose no threat to the possibility of voluntary action for which we can be held responsible, because there are still elements of the action performed in Libet's (and presumably Haynes's) experiment that are initiated by conscious mental events. For example, Owen Flanagan (1996) argues that, as long as taking part in Libet's experiment was consciously initiated, it does not matter if the realisation of this 'big picture' decision was not consciously initiated. It has also been suggested that as long as the precise details of the movement (e.g. whether to use one's left or right hand) are consciously initiated then it does not matter if the action as a whole was unconsciously initiated (Haggard and Libet 2001). However, these responses implicitly accept that an action needs to be initiated by a conscious mental event for it to count as free or voluntary or the kind of action that the agent would be responsible for.

Neil Levy (2005) challenges this assumption, arguing that the idea that only consciously initiated actions can be free is conceptually incoherent. Reaching a decision, Levy argues, cannot be something we consciously control. Deliberation may be a conscious activity but decisions themselves are not 'actions performed by consciousness', Levy claims; they are, rather, events we wait for and passively witness. Levy argues that decisions cannot be events we consciously control for the following reason:

> [D]ecision making is, or is an important element of, our control system, whereby we control our activity and thereby attempt to control our surroundings. If we were able to control our control system, we should require another, higher-order, control system whereby to exert that control. And if we had such a higher-order control system, the same problems would simply arise with regard to it. The demand that we exercise conscious will seems to be the demand that we control our controlling. And that demand cannot be fulfilled. (2005: 73)

I have a lot of sympathy with Levy's argument that there is something conceptually confused about the idea that decisions are conscious mental events which initiate voluntary actions. This is because I do not think it is obvious that the possibility of free or voluntary action depends on mental events being initiators of bodily movements. Those tempted to see Libet-style experiments as threatening to our conception of ourselves as capable of free action seem to presuppose free or voluntary action is only possible if mental items stand in causal relations to bodily movements.

Sophie Gibb also seems to endorse the idea that our conception of human agency entails that beliefs and desires stand in causal relations to bodily movements. She writes that:

> The thought that mental causes have physical effects—that our beliefs and desires can give rise to the movement of our bodies—is central to our pre-theoretical notion of human agency. (2013: 321)

What is interesting about Gibb's remark is that she also claims that this idea is a 'pre-theoretical notion'. This is a claim that I find surprising. Our everyday pre-theoretical way of talking about the mind and its place in the causal world does not usually involve talking about mental causes. We talk often about doing things because of what we believe, think, want or feel, but most of the time these discussions do not obviously involve identifying something that occurred at a particular time that triggered our action, or which moved us from a state of inaction to a state of action. As Elizabeth Anscombe noted, when 'one says what desire an act was meant to satisfy, one does not identify a feeling, image or idea that precedes the act the desire explains: one does not answer the question "what did you see or hear or feel, or what ideas or images cropped up in your

mind and led up to it?'" (2000: 17). This is not to say that we *never* speak about mental causes—of course we do. My claim is rather that a lot of talk about how our thoughts and feelings feature in our daily lives does not mention mental causes. There may be reasons to think that when we explain our actions in terms of our thoughts and feelings such explanations can only be true if mental causes have physical effects (I will examine those reasons in Chapter 2), but this would constitute a philosophical argument not a pre-theoretical notion. Indeed, it is precisely this philosophical argument that I believe is the reason the relational understanding of mental causation is so widely endorsed.

Within the philosophy of mind, and especially in discussions concerning physicalism, many believe that intentional or voluntary human action is possible only if mental items stand in causal relations to physical events such as bodily movements. The idea that psychological explanations of intentional action, and even the possibility of intentional action itself, entails the existence of causal relations between mental items and physical items is central to a view of intentional action that I call the 'causal theory of intentional action'. This theory of intentional action has become the dominant theory of intentional action. Although this theory concerns what makes an action intentional, and is logically independent from physicalism, I believe that implicit commitment to this theory is what makes the relational understanding of mental causation seem undeniable to physicalists. Physicalists, and those sympathetic to physicalism, assume the relational understanding of mental causation because of implicit acceptance of causal theories of intentional action. This is the first way in which different elements of the physicalist triad support each other.

1.4 Mental causation and the relational approach to causation

Physicalism and the relational approach to causation are also mutually reinforcing. This is because both views are thought to be *naturalistic*. Physicalism is often hailed as a naturalistic account of the mind, meaning that it is a metaphysics of mind that fits comfortably with a scientific view of the world, and especially a scientific view of causation. The thought is that physicalism does not commit us to the existence of anything that would be regarded as an unnatural addition to the world as described by science. Rightly or wrongly, the relational approach to causation is also part of this naturalistic worldview.

The connection between naturalism and the relational approach to causation is, I think, part of David Hume's influence on the philosophy of causation. Inspired by Hume's ideas about causation, philosophers of causation have often assumed that empirical science cannot provide us with any knowledge of necessitating connections in nature, whereby an object with certain powers 'must' behave in certain ways in certain conditions. Powers have long been regarded as epistemically suspicious and ineffable: we can perceive a thing's properties, what it is like, but not what it is capable of doing. Many philosophers of

causation therefore believe that causation, as it exists in reality, cannot be the exercise of power or efficacy or the making-happen of events if empirical science is to provide us with any knowledge of it. Hume himself tried to articulate a conception of causation divested of any association with natural necessity and did so by describing causation as a relation. Since then, philosophy of causation has proceeded under the assumption that there is one sort of thing that is causation: causation is a type of relation that holds independently of how the relata are described. To entertain the possibility that causation might be something richer than this, or that causation might refer to different things depending on the explanatory context, is to construe causation as something mysterious, ineffable and empirically unrespectable.

The mutually supportive connection between physicalism and the relational approach to causation can also be seen if we examine the principle of causal closure more closely. The principle of causal closure is a key part of any naturalistic account of the mind. If your account of the mind appeals to causal relations, and you want your account to be naturalistic, then the principle of causal closure is a constraint on what those causal relations can be like. The principle of causal closure is a key part of naturalism because it is supposed to be derived from, and supported by, the findings of scientific investigations into causal processes. Scientific discoveries are supposed to show that going beyond the physical realm to causally account for physical effects is unnecessary. The fact that physical science has been able to posit physical causes of many physical effects, and has never needed to 'leave the realm of the physical to find a fully sufficient cause' (Papineau 2001: 8), constitutes inductive evidence that every physical effect has a physical cause.

However, despite being readily assented to, the exact content of the causal closure principle is not obvious, as it is not clear what is meant by the terms 'physical' and 'sufficient cause'. Does it, for instance, mean that no physical effect can have as its cause a 'supernatural' substance? Or that no physical effect can have as its cause an event of a type that is not part of the subject matter of a physical science? These two interpretations are very different.

The first takes the referent of 'sufficient cause' to be a substance and 'physical' to mean anything that is not 'supernatural'. This version of the causal closure principle speaks against Cartesian minds being the causes of physical effects, it also seems to be empirically supported, but it does not immediately rule out that there might be physical effects that have non-physical *events* as causes. As Barry Stroud points out, the existential claim that the world contains only non-supernatural substances allows the character of the world to be 'as rich as you please' (1986: 264). Including irreducible, causally efficacious mental states and events is not at all problematic as long as they are states of, or events involving, non-supernatural beings.

The second interpretation takes the referent of 'sufficient cause' to be an event and 'physical' to mean strictly expressible in the vocabulary of physical science, which is a stronger thesis and more useful to the physicalist argument.

However, it is more difficult to justify empirically. A valid generalisation from the causal processes science investigates to all causal processes depends on the causal processes science investigates being suitably similar to other causal processes, including those that are supposed to involve mental causation. As we saw in the previous section, most arguments for physicalism focus on the mental causation associated with human action. However, are human actions suitably similar to the causal processes physical science investigates? Human actions are sometimes directed towards a goal, they can succeed or fail, they can be paused and then continued, and they can be started but never completed. In many ways, human actions are unlike the causal processes investigated by physical science; therefore, it is at least questionable whether a principle like causal closure should apply to them. However, this line of thought is unlikely to be persuasive if you are sympathetic to a relational approach to causation. According to the relational approach to causation, all causation is the same; causation is a homogenous phenomenon. Causation can hold between diverse relata, but there is one relation of causation, and that is all causation is. On this approach, then, the features I just ascribed to human actions should be analysable in terms of an event-causal sequence. They represent ways in which the causality of human action might involve unique relata (neural events maybe)—they do not represent unique features of the causation itself. Therefore, features such as goal-directedness do not represent properties that radically distinguish human actions from the kinds of causal processes investigated by physical science. All causation is the same; only the relata change. In this way, the relational approach to causation supports physicalism.

1.5 Troubles with physicalism

Physicalism has dominated philosophy of mind. It can seem like the only plausible option when faced with arguments centred around mental causation. However, I believe the reason for this dominance is because of physicalism's connection to causal theories of intentional action and relational approaches to causation and not because physicalism itself is the best, most informative metaphysics of the mind. As I said in the introduction, it is not my intention to argue directly against physicalism in this book. I do not intend to prove that physicalism must be false. This is because my main objection to the physicalist triad is not that physicalism itself is implausible or internally inconsistent. What I object to is physicalism's *hegemony*. Physicalism often seems like the only option. Debates can be had about what kind of physicalist you want to be, but the causal argument seems to make it impossible to doubt that some form of physicalism must be true. This is where my discomfort lies, as I believe the arguments for physicalism, the arguments that make physicalism seem like such an indubitable theory, are artificially bolstered by their association with causal theories of intentional action and relational approaches to causation.

Physicalism can only be given a fair assessment after the conception of mental causation assumed by arguments for physicalism—the conception of mental causation suggested by causal theories of intentional action and relational approaches to causation—is examined. Nevertheless, it is useful to say something about why someone might be dissatisfied with physicalism as a metaphysics of mind, as this will help justify my project.

First, there are those who doubt the coherence of physicalism. Physicalism says everything that exists is either itself a physical entity or 'nothing over and above' a physical entity—but what exactly is a physical entity? For some, this interpretive difficulty represents a fundamental flaw with physicalism as a metaphysical theory. It shows either that physicalism must be false or that it does not make a substantive claim.

In everyday discourse, we might take physical to mean 'relating to things we can sense' or 'extended in space' or 'not supernatural'. This everyday conception, however, is not precise enough to be useful in a formulation of physicalism. To say that everything that exists is related to things we can sense is very vague—it could potentially include mental states, properties and events, as these are related to human beings, and those are the kind of entity the physicalist wants to say do not fundamentally exist. To say that everything that exists is extended in space makes physicalism indistinguishable from materialism, the view that everything that exists is matter or made of matter. Materialism is no longer a popular view among philosophers as modern physics seems to countenance the existence of entities that are not matter or made of matter—photons, for example. To say that everything that exists is not supernatural rules out the existence of souls or purely mental substances but leaves room for the existence of irreducibly mental states, properties and events as long as they are states and properties of, or events involving, non-supernatural beings.

For many physicalists, specifying what it is to be physical must include reference to physical science. One way to do this is to define the physical entities as those which physical sciences tell us about. However, now we face the problem of deciding which sciences count as physical sciences. Physics obviously counts. Does biology? Does economics? Does psychology? There must be some way of drawing a boundary between physical sciences and non-physical sciences, otherwise the claim that everything is physical will be trivially true. One possibility is that physical sciences are those whose laws could be expressed in the vocabulary of *physics* with the help of suitable translation principles. Some sciences could not be reduced in this way—those sciences are the non-physical sciences. However, there is a famous issue with this suggestion. Carl Hempel (1969) pointed out that, if 'physical' is defined in terms of contemporary physics, then physicalism must be false as we cannot presume that contemporary physics is complete and describes everything that exists. However, if 'physical' is defined in terms of a future, complete physics, then physicalism is trivially true, because for all we know a future physics might include entities that the physicalist now wants to class as non-physical, namely mental entities. This is Hempel's dilemma.

In response to Hempel's dilemma, physicalists have suggested several other ways to define what 'physical' means. One way is to define physical in terms of what it is not: 'Physical entities are not fundamentally mental (that is, do not individually possess or bestow mentality)' (Wilson 2006: 61). This method of defining the physical allows physicalists to continue to use physics as the authority when it comes to describing the fundamental ontology of the world, acknowledging that current physics is incomplete and therefore that the world may contain more than current physics says it contains. However, the method avoids making physicalism trivially true because, although we do not know exactly what the ontology of a complete physics will be, we can say that it won't include fundamentally mental entities because fundamentally mental entities are not physical. Physicalism is a theory about what fundamentally exists and, as Papineau puts it, to do that 'it isn't crucial that you know exactly what a complete physics would include. Much more important is to know what it won't include' (2001: 12).

Another way to define 'physical' is inspired by structural realism, the view that we should believe to be true the structural content of our best scientific theories. According to this view, physical entities are those which are described by the mathematical language characteristic of physics. This means that the as-yet-undiscovered entities of future physics will count as physical because they will also be describable in this mathematical language (thus Hempel's dilemma is avoided) (Chalmers 2020).

Alyssa Ney (2008) further suggests that the best way to understand the content of physicalism is not to understand it as a truth-apt statement about what exists (and so not as the kind of statement that could be subject to Hempel's dilemma) but to see it as an 'oath' to formulate one's ontology according to physics. At least for now, physics does not include fundamentally mental entities; therefore, physicalists do not believe in fundamentally mental entities.

This is not an exhaustive list of all the ways physicalists have tried to solve the problem of specifying what is 'physical'. Suffice it to say that the debate is ongoing. For some, this issue is unsolvable. Tim Crane and D. H. Mellor argue that 'physicalism lacks a clear and credible definition, and that in no non-vacuous interpretation is it true' (1990: 394). A key premise of their argument is that, for physicalism to stand a chance of being non-vacuously true, there must be a reason why physical science is a 'unique ontological authority' (1990: 394). In other words, there must be a credible answer to the question 'what is special about physical science such that it can tell us what exists, whereas non-physical sciences cannot?' Crane and Mellor are of the opinion that there is no credible answer to this question.

I will not give a detailed assessment of Crane and Mellor's argument, but there are two interesting points they make in their 1990 paper that I want to highlight. The first is that physics might seem like an ontological authority because it can seem *universal*, i.e. about everything, and *basic*, i.e. about the most fundamental parts of the world. Crane and Mellor suggest that physics seems universal and basic because of its conventional subject matter. For example,

physics encompasses the study of the fundamental particles that compose all paradigmatically physical objects from rocks to animals to planets. Physics also encompasses mechanics and so studies everything that moves. A science with such vast reach can seem *universal*, and a science that studies the microscopic building blocks of the world can seem *basic*. However, just because a science has a far-reaching subject matter does not mean everything falls within its domain and just because a science studies the microscopic building blocks of a great many things does not mean it studies everything that exists.

The second point of interest is Crane and Mellor's argument against the suggestion that physical sciences enjoy ontological authority because only physical sciences describe causal reality. Crane and Mellor's view here is particularly interesting given the connections I believe exist between physicalism and certain assumptions about the nature of causation. Crane and Mellor argue that there is no good reason to deny that non-physical sciences like psychology describe causal reality. One bad reason (in their opinion) is the stipulation that for a causal claim to be true it must depend on an extensional causal relation between particulars. Crane and Mellor argue that there are examples within physics that show that this stipulation is false, and so there is no reason to deny that true causal claims can be made within psychology even though these claims often do not rest on an extensional causal relation between particulars. This discussion is interesting because it is another example of how physicalism derives support from substantive assumptions about the nature of causation. When those assumptions are questioned, the case for being a physicalist is weakened. It is also interesting because in Chapter 7 I will make a similar argument about the nature of causal explanations. I will argue that it is not necessary for an explanation to be causal that its explanandum designate an effect and its explanans designate an item which is the cause of that effect. In so doing, I will allow causal explanations that use mental concepts to count as revealing causal truths even though they do not designate particulars which are causes and particulars which are effects.

There are many options open to the physicalist when it comes to specifying what 'physical' means. It would be very presumptive of me to say none of those options is satisfactory; I cannot think of any *a priori* reason why one of the many options cannot be made to work. However, I think this interpretive difficulty shows that physicalism's credentials as an informative metaphysical theory are not as secure as they may first appear. A lot depends on how physicalism's core thesis is made precise.

A second major criticism of physicalism is that there are questions about the mind that physicalism cannot answer and for this reason it is an unsatisfactory account of mentality. The most famous version of this critique is David Chalmers's (1996) so-called 'hard problem of consciousness'. Chalmers's argumentative strategy is as follows. There are strong reasons for thinking that it is impossible to explain consciousness in physical terms. The idea that one day a scientist could

look at a person (or maybe their brain) and say, "they are having this conscious experience because they have these physical properties" and that constitute a fully satisfactory explanation is not credible (at least, according to Chalmers). Even if we imagine that every physical fact about a person is fully specified, it seems like we could still wonder why that person is having the conscious experience they are having. There is an 'explanatory gap' (Levine 1983). From this explanatory gap, Chalmers concludes that there is also a metaphysical gap. The reason we cannot explain how consciousness arises from physical states is because it does not arise from physical states. Physicalism can never be true of mental states that have a phenomenal quality, i.e. states that we consciously experience.

A vast literature exists on this topic. There are those who argue that drawing a metaphysical conclusion from an explanatory problem is unjustified (Hill 1997; Yablo 1993), there are those who argue that the explanatory gap problem is misconceived (Tye 1999) or that it is not as insurmountable as it seems (Dennett 2005), and there are those who argue that an explanatory gap is to be expected if physicalism is true and so actually confirms rather than disproves physicalism (Papineau 2002). I do not intend to get into this debate here. I bring up the issue of the hard problem to make a simpler point. The mind–body problem —the central issue in philosophy of mind—concerns the question of how physical things like human beings are capable of thought (including conscious thought). The explanatory gap seems to indicate that the mind–body problem is not satisfactorily answered by physicalism. Physicalism sidesteps the mind–body problem rather than directly confronting the issue of how our physical construction enables us to engage in mental activities like thinking, imagining and feeling. The fact that brain activity seems to have some connection to our ability to think is a genuinely puzzling fact—brain activity and thinking seem totally unlike each other. Physicalism seems to ignore rather than engage with that puzzlement. For this reason, there is an incentive to at least consider other options. So, even if Chalmers's argument against physicalism does not actually prove that physicalism is false, the hard problem gives us reason to consider alternative metaphysical accounts of the mind.

I have outlined some criticisms of physicalism discussed in philosophy of mind. I have not adjudicated on any of these issues. I think the problems outlined above are serious but it would take a much more detailed analysis to assess whether the problems are insurmountable and therefore falsify physicalism. Nevertheless, the foregoing discussion shows that physicalism is not without its critics. It is not a perfect metaphysics of mind; therefore, the fact that it can often seem like the only option is a problem. I believe that the physicalist triad has limited our thinking about mental causation and therefore prevented us from exploring more diverse accounts of the relationship between our mind and body. The following chapters investigate whether there is a way to break out of this triad, and thereby open up new ways of understanding mental causation. It is my hope that doing this will refresh debates in philosophy of mind

by allowing us to postulate new metaphysical theories of mentality that may be able to answer questions about the mind that physicalism cannot.

References[7]

Anscombe, G E M 2000 *Intention*. Cambridge, MA: Harvard University Press. Originally published in England in 1957 by Basil Blackwell.

Armstrong, D M 1968 *A materialist theory of the mind*. London: Routledge.

Árnadóttir, S T and Crane, T 2013 There is no exclusion problem. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 248–266.

Bennett, K 2003 Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs*, 37(3): 471–497. DOI: https://doi.org/10.1111/1468-0068.00447

Bryant, A 2020 Physicalism. In: Raven, M J *The Routledge handbook of metaphysical grounding*. New York: Routledge. pp. 484–500.

Burge, T 1979 Individualism and the mental. *Midwest Studies in Philosophy*, 4(1): 73–122. DOI: https://doi.org/10.1111/j.1475-4975.1979.tb00374.x

Chalmers, D 1996 *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Chalmers, D 2020 Is the hard problem of consciousness universal? *Journal of Consciousness Studies*, 27(5–6): 227–257.

Crane, T 1995 The mental causation debate. *Aristotelian Society Supplementary Volume*, 69(1): 211–236. DOI: https://doi.org/10.1093/aristoteliansupp/69.1.211

Crane, T and Mellor, D H 1990 There is no question of physicalism. *Mind*, 99(394): 185–206. DOI: https://doi.org/10.1093/mind/xcix.394.185

Davidson, D 1970 Mental events. In: Foster, L and Swanson, J W *Experience and theory*. 2nd ed. Cambridge, MA: University of Massachusetts Press. Reprinted in Davidson 2001 pp. 207–224.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Dennett, D 2005 *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.

Elpidorou, A and Dove, G 2018 *Consciousness and physicalism: A defense of a research program*. New York: Routledge.

Fifel, K 2018 Readiness potential and neuronal determinism: New insights on Libet experiment. *Journal of Neuroscience*, 38(4): 784–786. DOI: https://doi.org/10.1523/JNEUROSCI.3136-17.2017

[7] Author note: some references to Davidson are formatted (1963/2001). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001 collection of his essays, with the page numbers relating to that volume.

Flanagan, O J 1996 *Self expressions: Mind, morals, and the meaning of life*. Oxford: Oxford University Press.

Gibb, S 2013 Introduction to mental causation and ontology. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 1–17.

Haggard, P and Libet, B 2001 Conscious intention and brain activity. *Journal of Consciousness Studies*, 8(11): 47–63.

Haugeland, J 1982 Weak supervenience. *American Philosophical Quarterly*, 19(1): 93–103. http://www.jstor.org/stable/20013945

Haynes, J-D and Pauen, M 2013 The complex network of intentions. In Caruso, G D *Exploring the illusion of free will and moral responsibility*. Plymouth: Lexington Books. pp. 221–238.

Heil, J 2013 Mental causation. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 18–35.

Hellman, G P and Thompson, F W 1975 Physicalism: Ontology, determination, and reduction. *Journal of Philosophy*, 72(17): 551–564. DOI: https://doi.org/10.2307/2025067

Hempel, C 1969 Reduction: Ontological and linguistic facets. In White, M et al. *Philosophy, science, and method: Essays in honor of Ernest Nagel*. New York: St Martin's Press. pp. 179–199.

Hill, C S 1997 Imaginability, conceivability, possibility and the mind-body problem. *Philosophical Studies*, 87(1): 61–85. DOI: https://doi.org/10.1023/a:1017911200883

Hopkins, J 1978 Mental states, natural kinds and psychophysical laws. *Aristotelian Society Supplementary Volume*, 52(1): 195–236. DOI: https://doi.org/10.1093/aristoteliansupp/52.1.195

Horgan, T E 1993 From supervenience to superdupervenience: Meeting the demands of a material world. *Mind*, 102(408): 555–586. DOI: https://doi.org/10.1093/mind/102.408.555

Hornsby, J 2015 Causality and 'the mental'. *HUMANA.MENTE Journal of Philosophical Studies*, 8(29): 125–140.

Jackson, F 1998 *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Kim, J 1989 The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63(3): 31–47. DOI: https://doi.org/10.2307/3130081

Kim, J 1993 *Supervenience and the mind: Selected essays*. Cambridge: Cambridge University Press.

Kim, J 1998 *Mind in a physical world*. Cambridge: Cambridge University Press.

Kim, J 2001 Mental causation and consciousness: The two mind-body problems for the physicalist. In: Carl, G and Loewer, B *Physicalism and its discontents*. Cambridge: Cambridge University Press. pp. 271–283.

Kim, J 2005 *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.

Kroedel, T 2020 *Mental causation: A counterfactual theory*. Cambridge: Cambridge University Press.

Levine, J 1983 Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(October): 354–361. DOI: https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Levy, N 2005 Libet's impossible demand. *Journal of Consciousness Studies*, 12(12): 67–76.

Libet, B 1985 Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4): 529–566. DOI: https://doi.org/10.1017/s0140525x00044903

Libet, B 1999 Do we have free will? *Journal of Consciousness Studies*, 6(8–9): 47–57.

Libet, B 2002 The timing of mental events: Libet's experimental findings and their implications. *Consciousness and Cognition*, 11(2): 291–299. 10.1006/ccog.2002.0568

Lowe, E J 2000 Causal closure principles and emergentism. *Philosophy*, 75(294): 571–586. DOI: https://doi.org/10.1017/s003181910000067x

List, C and Menzies, P 2009 Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy*, 106(9): 475–502. DOI: https://doi.org/10.5840/jphil2009106936

Lynch, M P and Glasgow, J 2003 The impossibility of superdupervenience. *Philosophical Studies*, 113(3): 201–221. DOI: https://doi.org/10.1023/a:1024037729994

Melnyk, A 2003 *A physicalist manifesto: Thoroughly modern materialism*. New York: Cambridge University Press.

Melnyk, A 2006 Realization and the formulation of physicalism. *Philosophical Studies*, 131(1): 127–155. DOI: https://doi.org/10.1007/s11098-005-5986-y

Melnyk, A 2018 In defense of a realization formulation of physicalism. *Topoi*, 37(3): 483–493. DOI: https://doi.org/10.1007/s11245-016-9404-1

Menzies, P 2013 Mental causation in the physical world. In: Gibb, S. C., Lowe, E. J., and Ingthorsson, R. D. *Mental causation and ontology*. Oxford: Oxford University Press. pp. 58–88.

Morris, K 2018 Truthmaking and the mysteries of emergence. In Vintiadis, E and Mekios, C *Brute facts*. Oxford: Oxford University Press. pp. 113–129.

Ney, A 2008 Defining physicalism. *Philosophy Compass*, 3(5): 1033–1048. DOI: https://doi.org/10.1111/j.1747-9991.2008.00163.x

Noordhof, P 1999 The overdetermination argument versus the cause-and-essence principle—no contest. *Mind*, 8(430): 367–375. DOI: https://doi.org/10.1093/mind/108.430.367

Papineau, D 1993 *Philosophical naturalism*. Oxford: Blackwell.

Papineau, D 2001 The rise of physicalism. In: Gillett, C and Barry, L *Physicalism and its discontents*. Cambridge: Cambridge University Press. pp. 3–26.

Papineau, D 2002 *Thinking about consciousness*. New York: Oxford University Press.

Pettit, P 1993 A definition of physicalism. *Analysis*, 53(4): 213–223. DOI: https://doi.org/10.1093/analys/53.4.213

Putnam, H 1975 The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7: 131–193.

Ryle, G 1949 *The concept of mind*. London: Hutchinson's University Library.

Shoemaker, S 2001 Realization and mental causation. In Gillett, C and Loewer, *The proceedings of the Twentieth World Congress of Philosophy*. Cambridge: Cambridge University Press. pp. 23–33.

Shoemaker, S 2007 *Physical realization*. Oxford: Oxford University Press.

Shoemaker, S 2013 Physical realization without preemption. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 35–57.

Skillen, A 1984 Mind and matter: A problem which refuses dissolution. *Mind*, 93(372): 514–526. DOI: https://doi.org/10.1093/mind/xciii.372.514

Steward, H 1997 *The ontology of mind: Events, processes and states*. Oxford: Oxford University Press.

Stroud, B 1986 The physical world. *Proceedings of the Aristotelian Society*, 87(1): 263–277. DOI: https://doi.org/10.1093/aristotelian/87.1.263

Sturgeon, S 1998 Physicalism and overdetermination. *Mind*, 107(426): 411–432. DOI: https://doi.org/10.1093/mind/107.426.411

Tye, M 1999 Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind*, 108(432): 705–725. DOI: https://doi.org/10.1093/mind/108.432.705

White, A 2020 Processes and the philosophy of action. *Philosophical Explorations*, 23(2): 112–129. DOI: https://doi.org/10.1080/13869795.2020.1753801

Wilson, J 2005 Supervenience-based formulations of physicalism. *Noûs*, 39(3): 426–459. DOI: https://doi.org/10.1111/j.0029-4624.2005.00508.x

Wilson, J 2006 On characterizing the physical. *Philosophical Studies*, 131(1): 61–99. DOI: https://doi.org/10.1007/s11098-006-5984-8

Wilson, J 2016 Grounding-based formulations of physicalism. *Topoi*, 37(3): 495–512. DOI: https://doi.org/10.1007/s11245-016-9435-7

Yablo, S 1992 Mental causation. *Philosophical Review*, 101(2): 245–280. DOI: https://doi.org/10.2307/2185535

Yablo, S 1993 Is conceivability a guide to possibility? *Philosophy and Phenomenological Research*, 53(1): 1–42. DOI: https://doi.org/10.2307/210805

CHAPTER 2

Causal Theories of Intentional Action

In this chapter I turn my attention to the second element of the physicalist triad: causal theories of intentional action. Two central questions within philosophy of action are 'how do reasons explain actions?' and 'what is the nature of intentional action?'. The two questions are related, as part of what makes intentional actions distinctive is that often (but not always) when we explain an intentional action, that is, say why the agent acted as she did, we do so by giving the agent's reason for acting as she did.[8] Explanations that cite an agent's reasons are called 'rationalising explanations'. Rationalising explanations explain why an agent acted as she did (this is the explanandum) by telling us why, in the agent's eyes, what she did was a rational thing for her to do (this is the explanans). The nature of intentional action is thus inseparable from intentional action's appropriateness for receiving rationalising explanations. Whatever intentional actions are, they must be things that can be explained by reasons.

The first question concerns *how* rationalising explanations explain. How does a statement telling us why what an agent did seemed to them to be rational explain why the agent did as she did? How does the explanans of a rationalising explanation illuminate the explanandum? An influential answer to this question is the answer offered by Donald Davidson. Davidson (1963) argued that rationalising explanations are causal explanations. Davidson claimed that the explanantia of rationalising explanations are facts about what the agent wants to do (or what the agent has an urge to do, or what the agent has an ambition to do) and facts about what the agent believes about how to do it. Davidson calls

[8] Two examples of an explanation of an intentional action that do *not* cite the agent's reasons or motives are: 'Sally bit the policeman because she was drunk' (Hyman 2015: 105) and 'She threw the water at him because she was angry at him'.

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 39–54. York: White Rose University Press. DOI: https://doi.org/10.22599/White.c.

the composite of a desire to perform some type of action and a belief about how performance of that action may be achieved 'the *primary reason* why the agent performed the action' (1963/2001: 4, emphasis in original). Davidson argued that, when we say the agent acted as she did *because* she wanted to do something, or *because* she believed something was the case, this 'because' implies causality. From this, Davidson concluded that states of desiring and states of believing—or, at least, events suitably related to states of desiring and states of believing, such as the onset of the desire or the onset of the belief—are causes of the actions they explain. Davidson's view is commonly called the *causal theory of action explanation*.

The second question concerns what intentional actions are. One answer is that they are events, and basic actions (i.e. actions not done *by* doing something else) are bodily movements. For example, the action of raising my arm is one and the same event as my arm's rising (Davidson 1987: 37). This is not yet a complete answer, as not all bodily movements are intentional actions. Epileptic fits are bodily movements but they are not intentional. To complete the story, several philosophers have suggested that bodily movements count as intentional actions when and only when they are caused, in the right way, by mental states of the agent that also rationalise the action (e.g. Bishop 1989; Davidson 1963; Davidson 1971; Mele 2003; Smith 2012). This answer has become the standard account of intentional action and is commonly called *the causal theory of action*.

In the previous chapter I argued that the relational understanding of mental causation, which plays a pivotal role in arguments for physicalism, is made to seem indispensable because of implicit acceptance of these causal theories of intentional action. Many physicalists believe that intentional or voluntary human action is only possible if mental items stand in causal relations to physical events such as bodily movements. In Sections 2.1 and 2.2 I will explain in more detail how causal theories of intentional action reinforce physicalism about mentality, which will help us see that the best strategy for resisting physicalism about mentality will involve challenging key aspects of the causal theories of intentional action. In Section 2.3 I will explain how causal theories of intentional action are themselves supported by relational assumptions about the nature of causation.

2.1 Rationalising explanations, mental concepts and mental causation

On the causal theory of action explanation, rationalising explanations explain by giving a causal account of the agent's action. That is, a statement telling us why what an agent did seemed to them to be a rational thing to do explains why the agent did as she did by giving us causal information. Davidson's (1963) argument for this position is best thought of as a challenge to anyone who thinks

that rationalising explanations are not causal, as Davidson does not offer any positive reason to think that they are.

In brief, Davidson's argument is as follows. *Some* statements that tell us why what an agent did seemed to them to be rational do *not* explain why the agent did as she did. This kind of statement could be called a 'mere rationalisation'. Mere rationalisations are similar to rationalising explanations in that they also tell us why the course of action taken by the agent seemed, to the agent, to be a rational course of action to take. However, mere rationalisations do not tell us *why an agent acted as she did*—they *only* tell us why what the agent did seemed, to the agent, to be a rational thing for them to do. For example, imagine that Anna is deciding whether or not to speak at a conference. She knows that speaking at a conference will be good for her career, but in the end, she decides to speak at the conference because it will draw praise from her friends, and not because it will be good for her career (perhaps she does not really care about her career). Anna actually spoke at the conference because she would get praise from her friends, not because it would be good for her career. In this context, the following statement would be a mere rationalisation of Anna's action:

(a) Speaking at the conference seemed rational to Anna because it would be good for her career.

This is a mere rationalisation because it explains why speaking at the conference seemed to Anna to be a rational thing for her to do—but it does not explain why Anna actually spoke at the conference. It is not true that Anna spoke at the conference *because* she thought it would help her career. On the other hand, it is true that Anna spoke at the conference because she would receive praise from her friends. That Anna would receive praise from her friends if she spoke at the conference *does* explain why Anna acted as she did. Because some statements which tell us why what an agent did seemed to them to be rational do not explain why the agent did as she did, those statements that do both must achieve this by doing more than simply revealing why what an agent did seemed to them to be a rational thing to do. And if the extra thing rationalising explanations do is not revealing causal information, what is it? This question has come to be known as 'Davidson's challenge' and Davidson thinks there is no satisfactory answer to it.

Jonathan Dancy (2000) denies that successful rationalising explanations do more than reveal why what an agent did seemed to them to be a rational thing to do. The difference between statements that rationalise but do not explain and statements that rationalise *and* explain is simply that, in the former, the belief/desire mentioned is not the belief/desire the agent acted in the light of, and in the latter the belief/desire mentioned *is* the belief/desire the agent acted in the light of. Davidson insists that the explanatory connection between beliefs/desires an agent acts in light of and the agent's action cannot be primitive—it has to hold in virtue of some other connection between the agent's beliefs/

desires and their action. But, Dancy objects, Davidson provides no argument against the following view:

> [T]he difference between those reasons for which the agent did in fact act and those for which he might have acted but did not is not a difference in causal role at all. It is just the difference between the considerations in the light of which he acted and other considerations he took to favour acting as he did but which were not in fact ones in the light of which he decided to do what he did. (2000: 163)

In other words, Dancy doesn't think that Davidson provides any argument against taking 'acted in the light of' as primitive.

On Dancy's view, 'acted in the light of' performs the function in the case of rationalising explanations that truth plays in the case of other sorts of explanation. Like truth, 'acted in the light of' is a status capable of belonging to statements given as explanans, which is a necessary condition for their explanatoriness. For example, compare 'George is the firstborn of William and Kate' with 'George is the firstborn of Elizabeth and Philip' as putative explanans of the following explanandum: why is George heir to the throne? Both statements posit the kind of relationship that would guarantee George's being the heir to the throne, but only the first statement can genuinely explain why George is heir to the throne because only the first statement is true. There is nothing perplexing about the fact that truth can make the difference between two statements that both posit something that would make sense of the explanans. That only true statements can explain is plausibly a brute fact.

However, I think there *is* something perplexing about the fact that 'acted in the light of' also seems to be able to perform this function. That 'acted in the light of' can perform this function seems like something that needs accounting for—it does not seem like a brute fact. There must be something about statements that tell us the reason the agent 'acted in the light of' that grounds their explanatoriness. The question Davidson's challenge raises is: *why* does learning that Anna's reason for acting was that she would receive praise *explain* why Anna spoke at the conference? Why does 'acted in the light of' bestow explanatory power? Julia Tanney (2009) expresses the puzzle well:

> Davidson claims that it would be a mistake to conclude from the fact that placing the action in a larger pattern explains it, we now understand the sort of explanation involved, and that 'cause and effect form the sort of pattern that explain the effect in the sense of "explain" that we understand as well as any' [(1963/2001: 10)]. Davidson challenges the opponents of the causal view to identify what other pattern of explanation illustrates the relation between reason and action if they wish to sustain the claim that the pattern is not one of cause and effect. (2009: 96)

The task is to spell out what 'pattern of explanation' is demonstrated by rationalising explanations.

I have said that Davidson thought that the pattern of explanation demonstrated by rationalising explanations is a causal one. That is, that rationalising explanations explain by giving a *causal* account of the agent's action. However, what is the nature of the causal information rationalising explanations are supposed to provide? This question is particularly important as it has a bearing on how we ought to understand mental causation.

Davidson's answer is that 'the primary reason for an action is its cause' (1963/2001: 4). It is worth taking some time to explain what Davidson means by this. In Davidson's view, the explanantia of rationalising explanations are facts about what the agent wants to do and facts about what the agent believes about how to do it. Davidson calls the dual possession of a desire to perform some type of action and a belief about how performance of that action may be achieved 'the *primary reason* why the agent performed the action' (1963/2001: 4, emphasis in original). Davidson argued that 'For us to understand how a reason of any kind rationalises an action it is necessary and sufficient that we see, at least in essential outline, how to construct a primary reason' (1963/2001: 4).

I think that Davidson is essentially correct on this first point. I assume that explanation is a relation between facts and only facts can explain other facts.[9] Furthermore, I agree that the explanatory power of rationalising explanations rests on our ability to identify facts about an agent's desires and beliefs from the statement that rationalises the agent's action. Of course, rationalising explanations do not typically take the form 'agent A φed because A wanted to φ and believed that ψing was a way to φ'. Sometimes this is because it suffices to explain why someone acted as they did to only mention what the agent wanted to do. For example, in (b) Beth's action is explained in terms of her desire only:

(b) Beth is buying flour because she wants to make bread.

We do not need to be told that Beth believes or knows that buying flour is an essential preparatory action for making bread. We take it for granted that Beth possesses this knowledge.

Other times it is sufficient only to mention what the agent believes, or knows, about how to achieve what they want to do. For example, in (c), Carlin's action is explained in terms of his belief only:

[9] Van Fraassen (1980: 134–153) proposes a theory of explanations as answers to why-questions where both the answer and the topic of the why-question are true propositions. Raley (2007) has also defended the view that all explanation is factive. See also: Bokulich (2011), Hempel and Oppenheim (1948), Kitcher (1989) and Woodward (2003).

(c) Carlin is adding rosemary to the sauce because he believes it will make it taste better.

We do not need to be told that Carlin wants to make the sauce taste better—we take it for granted that he wants this. Davidson's point is not that all rationalising explanations *explicitly* give the primary reason why the agent acts but rather that, for the explanans of a rationalising explanation to illuminate the explanandum, 'it is necessary and sufficient that we see, at least in essential outline, how to construct a primary reason' (1963/2001: 4). That is, the explanatory power of rationalising explanations rests on our ability to construct a primary reason from any rationalising explanation.

Although I think Davidson is broadly correct in thinking that the explanantia of rationalising explanations are facts about what the agent wants and believes, there is a complication. When an agent acts for a reason, the reason for which they act is not usually a fact about the agent's own mental states. For example:

(d) Daniel took the A road because the motorway was shut.

In (d) Daniel's reason is 'that the motorway was shut', not 'that Daniel believed or knew that the motorway was shut'. At least, that is how things seem. How does this square with Davidson's claim that the primary reason why an agent acts is a belief–desire pair? The best way to tackle this complication is, I think, to acknowledge that the word 'reason' can be used in more than one way.

First, the term can be used to denote an agent's *reason for acting*. I follow Maria Alvarez (2010) in thinking that an agent's reason for acting is that which makes the action a sensible or rational or good thing to do. As Alvarez puts it, an agent's reason for acting is 'the desirability characterisation' the action has for the agent. As such, reasons for acting are not usually facts about an agent's mental states. Strictly speaking, Daniel's reason for taking the A road is not that he *wants* to get somewhere and *believes* that, because the motorway is shut, taking the A road is the only means of getting there. The good Daniel sees in taking the A road is that, given that the motorway is shut, taking the A road is the only way he can get to where he wants to go.

As well as being used to denote the desirability characterisation an action has for an agent, the word 'reason' can also be used as a synonym for 'explanans'. When we give *the reason why* such and such is the case, we are providing an explanans. Reasons why are explanantia of explanations. I think Davidson's claim that primary reasons given by rationalising explanations are belief–desire pairs is plausible only if 'primary reason' is taken to mean 'primary reason why' or 'primary explanans', because reasons for acting are not usually facts about the agent's own mental states. However, I believe that primary reasons why, i.e. the primary explanantia, of rationalising explanations *are* facts about what the agent wants and believes. That is, I believe that the explanatory power of rationalising explanations rests on our ability to construct a belief–desire pair from any rationalising explanation.

We are now in a better position to clearly state what Davidson means by the claim 'the primary reason for an action is its cause' (1963/2001: 4). Davidson's view is not only that rationalising explanations give causal information but that rationalising explanations are true if and only if the belief or desire which explains the action (or some mental event suitably related to the belief or desire) *stands in a causal relation to the action explained*. Davidson is making *two* claims here. First, rationalising explanations give causal information. Second, rationalising explanations are true if and only if the belief or desire which explains the action stands in a causal relation to the action explained. If Davidson is correct, then the possibility of true rationalising explanations of action entails that there must be causal relations between mental items and actions.

Construing rationalising explanations as explanations which posit an entity that is causally related to the action explained encourages us to think that concepts like *belief* and *desire* refer to mental *items*. This view, I believe, legitimises a metaphysics of mind wherein our status as minded creatures depends on the existence of mental events and states whose nature we have yet to discover and whose existence must, one way or another, be reconciled with the idea that the world is physical in all its fundamental aspects. In this way, the causal theory of action explanation creates the problem physicalism is supposed to solve. The causal theory of action explanation encourages us to accept an ontology that includes mental items whose intrinsic nature is up for discovery, which stand in causal relations to human actions. If we also assume that human actions fall under the jurisdiction of scientific causal explanation, then, unless the intrinsic nature of those mental items is, somehow, exhaustively determined by the underlying physical causes of our actions, it is hard to see how rationalising explanations can be true. To put it another way, if the causal theory of action explanation is correct, then the possibility of true rationalising explanations of action entails that there must be causal relations between mental items and actions. If we also assume that actions are physical events, then the causal theory of action explanation justifies the relational understanding of mental causation, which says that mental items stand in causal relations to physical events. And, as we saw in the previous chapter, the relational understanding of mental causation is the driving force in arguments for physicalism.

2.2 The causal theory of action and physicalism

The causal theory of action concerns the ontological question 'what is the nature of intentional action?' Although I have introduced the causal theory of action as if it were one unified theory, in fact matters are more complicated than this. There are many different theories that attempt to give a causal account of intentional action. What these many theories have in common is the commitment that acting intentionally consists in events being caused to happen by non-actional mental antecedents. However, there is plenty of room for disagreement after this commitment is accepted.

Most causal theorists believe that actions, or at least basic actions (i.e. actions we perform without having to do anything else first) are bodily movements. However, some causal theorists believe that actions are composite events such as the event of an-intention-causing-a-bodily-movement, or an event that involves neural states and bodily movements. For example, Michael Smith argues that 'we should suppose that actions are events that begin in the brain, continue on in the nervous system and muscles, and end with the relevant events of the body's moving' (2021: 7).

There is also disagreement on exactly what kind of mental antecedents must cause an event to happen if it is to count as an intentional action. Following Davidson's suggestion that 'the primary reason for an action is its cause' (1963/2001: 4), some causal theorists have suggested that beliefs and desires must feature in the aetiology of an event, if that event is to count as an intentional action. For example, elsewhere Smith has proposed that:

> [A]ctions are those bodily movements that are caused and rationalised by a pair of mental states: a desire for some end, where ends can be thought of as ways the world could be, and a belief that something the agent can just do, namely move her body in the way to be explained, has some suitable chance of making the world the relevant way. Bodily movements that occur otherwise aren't actions, they are mere happenings. (2004: 165)

Some causal theorists take the mental antecedent necessary for intentional action to be an intention. On this kind of view, the agent's beliefs and desires cause the acquisition of an intention to act, which in turn triggers the behaviour that constitutes the agent's action. John Searle (1983) argues that, for an event to count as an action, the event must be caused by a specific kind of intention, namely one that continues exerting causal influence over an agent's behaviour even after the behaviour has begun, thereby sustaining and guiding the behaviour to ensure that it satisfies the agent's prior motive. Berent Enç (2003) also argues that for an event E to be an action it must be caused (in the way it is normally caused) by an intention, the content of which explicitly refers to bringing about an E-type event (2003: 78–79). For Enç, deliberation about what to do is a 'computational process … the causal consequence of which is the formation of an intention' that in turn causes a 'behavioural output' (2003: 2). Others suggest that second-order desires, like the desire to act on a particular motive (Frankfurt 1978) or the desire to act in accordance with reasons (Velleman 1992), must be part of the causal history of an event if that event is to count as an intentional action. In all these versions of the causal theory of action, mental items are assigned a causal role in bringing about an event.

Perhaps the most significant source of disagreement concerns what constitutes *the right way* for a mental item to cause a bodily movement for there to

be intentional action. Not just any causal chain from mental event to physical event is sufficient for there to be an intentional action. A necessary condition for acting intentionally is that the agent is in control of what is going on with them. It is difficult to explain what is meant by control in this context without begging the question against certain theories of action. I will have more to say on what kind of control is necessary for intentional action in later chapters, but for now it suffices to illustrate with an example what kind of control is required for intentional action.

Imagine my friend Amy really wants me to make tea, so she makes sure I am thirsty by giving me something salty to eat, puts a cup and some teabags nicely in view, then says, "Why don't you have some tea?" The conditions are right for me to make tea, but whether or not I do is still up to me. I am in control of my making tea (or not) in this case. Now, suppose Amy instals some clever machinery to manipulate my brain and nervous system and uses that to make me make tea (in the manner of the character Black from Harry Frankfurt's (1969) thought experiment). In this case, I am not in control of my movements. Amy has taken control over what goes on with me.

The causal theorist would say that the difference between these two cases, what explains why I have control in the one case but not in the other, has something to do with the causal history of my movements in each case. In the second case, where Amy manipulates my brain, the causal chain leading up to my bodily movement is not the kind of causal chain required for there to be agential control. For one thing, the causal chain does not involve my own mental states. However, it is not sufficient merely to include mental states in the causal chain leading to bodily movement. These mental states have to operate in the causal chain in the right way. For there to be intentional action, the causal chain from mental item to bodily movement must be such that it constitutes the agent's control over their action. The causal chain cannot deviate from the kind of causal chain that occurs in a normal, uncontroversial case of intentional action. Davidson gives an example of a deviant kind of causal chain:

> 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, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally. (Davidson 1973/2001: 79)

In this example, the climber has an end he wants to achieve and a belief about how to achieve this end. This belief–desire pair causes a bodily movement of a type that is rationalised by the belief–desire pair, just as causal theorists allege it would in an ordinary case of intentional action. But, in this case, the climber did not let go intentionally. There is great disagreement on what kind of causal

chain from mental state to bodily movement is required for the agent to retain control over their action. I will discuss this problem, known as the problem of deviant causal chains, further in Chapter 4.

Finally, there is disagreement on what exactly the causal theory of action should be a theory of. Some versions of the causal theory of action are presented as accounts of agency in general. These are presented as theories that explain the difference between things that you do and things that befall you, or 'between bodily movements that you are making happen and those which happen without your making them occur' (Brent 2017: 656), or 'between actions and things that we do when we are merely passive recipients of courses of events' (Enç 2003: 2). Other versions of the causal theory of action take for granted the distinction between events to which a person is subject and events of which the person is the agent, and offer specifically a theory of intentional action (Mele 1992; Mele 2003) or rational agency (Bratman 2001; Velleman 1992).

These disagreements are related to important questions about the nature of agency and intentional action. Some of these questions will arise again in later chapters, either when I critically evaluate causal theories of intentional action or when I present my own account of intentional action. For now, though, most of these disagreements can be set aside. To see the connection between causal theories of action and physicalism it is the core ontological commitment of all causal theories of intentional action that we need to focus on.

All versions of the causal theory of action hold that acting intentionally consists in the right kind of event being caused to happen, in the right way, by the right kind of mental antecedents. This commitment entails that acting intentionally is nothing over and above some special kind of event causation, and that the possibility of intentional action requires that certain mental items stand in causal relations. Just like the causal theory of action explanation, the causal theory of action encourages us to accept an ontology that includes mental items which stand in causal relations.

According to most versions of the causal theory of action, what mental items cause is either a physical event such as a bodily movement or an event that is composed or realised by a bodily movement. This means that, if the causal theory of action is correct, then the existence of causal relations between mental items and physical events (or events realised by physical events) is entailed by the existence of intentional action. If the causal theory of action is correct, then Kim's claim that 'the possibility of human agency … requires that our mental state have causal effects in the physical world' is also correct. In this way, the causal theory of action serves as justification for the relational understanding of mental causation.

It is difficult to endorse the causal theory of action without also being a physicalist, as the ontological component of the causal theory of action seems to set up the conditions for the causal argument for physicalism. Alfred Mele acknowledges the connection between taking a causal perspective on intentional action and physicalism. He states that the causal perspective 'is usually

embraced as part of a naturalistic stand on agency according to which mental items that play a causal/explanatory role in intentional conduct bear some important relation to physical states and events' (2003: 6). John Bishop also acknowledges this point:

> Surely we may understand how agency is naturally possible only if we first understand how mentality may be part of nature? That this is so is entirely clear if a Causal Theory of Action is to provide the solution to the problem of natural agency because this theory holds that action consists in behaviour caused by relevant mental states. And there is problem posterior to the problem of natural agency—namely, the problem of explaining how those extra properties beyond agency as such that are required for personal moral responsibility can themselves be realised within a natural scientific ontology. (1989: 8)

In Bishop's view, a complete naturalisation of our perspective of ourselves as agents capable of rational, intentional action would require a solution to 'scepticism about understanding how minds can be part of nature' (1989: 8). However, I think Bishop mischaracterises the connection between these two projects. He presents the problem of providing a naturalistic account of the mind as 'posterior' to the problem of finding a naturalistic account of agency. This implies that the former problem is in some ways independent from the latter problem. In my view, the connection between the project of naturalising the mind and the project of naturalising agency is much closer. It is the causal theory of action that encourages us to accept an ontology of causally efficacious mental items, an ontology that then needs to be reconciled with the 'naturalistic' view of what causation in the actual world is like. In other words, the causal theory of action justifies the relational understanding of mental causation, which as we have seen is the crucial premise in arguments for physicalism. For this reason, even though it is logically possible to accept the causal theory of action without being a physicalist, in practice belief in the causal theory of action supports physicalism. Because of this connection, the strongest challenge to causal arguments for physicalism will require a critical examination of the causal theory of intentional action.

I also think it is difficult to be a physicalist without endorsing the causal theory of action. This is because both theories are thought to be consistent with naturalism, a philosophical position that eschews the existence of anything that would be regarded as an unnatural addition to the world as described by science. Physicalism assumes nothing more than a world of physical things and this ontology is thought to fit comfortably with a scientific view of the world. Bishop argues that one should endorse the causal theory of action because it promises to 'make intelligible the possibility of agency within the natural order' (1989: 10). Thus, a key motivation for adopting causal theories of intentional action is that they seem to provide a naturalistic account of intentional action.

As Enç puts it, the causal theorist's starting point is that 'by assuming nothing more than a world of material things, we can understand the nature of decisions, of intentions, of voluntary action, and the difference between actions and things that we do when we are merely passive recipients of courses of events' (2003: 2). In this way, belief in physicalism lends credence to causal theories of intentional action, because both are apparently part of a naturalistic worldview. However, to see exactly *why* causal theories of intentional action are thought to be naturalistic it is necessary to examine the connections between causal theories of intentional action and the other element of the physicalist triad: relational approaches to causation.

2.3 Naturalistic agency and the relational approach to causation

The causal theory of action is reductive: it says that intentional action is nothing over and above event causation. The agent's role in bringing about what she intends is reduced to causation by her mental states or events. Agential control over what goes on exists, but it is exhaustively determined by some special kind of event causation. One key draw of causal theories of intentional action is that we can achieve an adequate understanding of intentional action without countenancing the existence of irreducible agent causation.

This is good, causal theorists argue, because the idea that there is the flux of causally related events and then *there are also* agents—three-dimensional substances, persons—who interfere with this flux to bring about the events they want to see happen is antithetical to the naturalistic view of the causal world. As we saw in the previous chapter, a naturalistic view of the causal world is one that endorses the relational approach to causation. According to naturalism, causation, as it exists in reality, cannot be the exercise of power because that kind of 'necessary connexion' is ineffable and empirically unrespectable. Instead, causation must be a certain kind of relation between events. The causal theory of action thus presupposes a metaphysics where causation is always, everywhere a relation between events. This approach to causation compels the causalist to seek to understand intentional action in terms of a distinction between different types of event causation. Causal reality is nothing more than a chain of causally related events, so, if intentional agency is a causal phenomenon at all, it must be located within this worldview. If you endorse the relational approach to causation, then the causation demonstrated in intentional action must be a relation, because all causation is, and will count as mental causation if and only if at least one of the terms of that relation is a mental entity.

The relational approach to causation is also presupposed by Davidson in his discussion of whether rationalising explanations of actions are causal explanations. Recall that Davidson argues that when we say the agent acted as she did

because she wanted to do something, or *because* she believed that something was the case, this 'because' implies causality. He also concludes from this that states of desiring and states of believing—or, at least, events suitably related to states of desiring and states of believing—are causes of the actions they explain. Davidson is assuming here that, if rationalising explanations reveal causal information, the causal information they reveal is that there are mental items, which the mental concepts employed in rationalising explanations pick out, that stand in causal relations to actions.

Contemporary non-causalists, who deny that rationalising explanations are causal explanations, also make this assumption. Julia Tanney is explicit about this:

> [T]he position I wish to bring back into focus says that what it is for an action to be in execution of an intention or for it to be explicable by reasons is not a matter of there being a causal relation [understood as 'a relation between two logically and temporally distinguishable events'] between intention or reasons and action. If causation is to be thus understood the pattern in virtue of which a person's intentions, motives or reasons explain her action is not *eo ipso* causal. (Tanney 2009: 95)

However, is it right to assume that a rationalising explanation is causal only if it posits a causal relation between an item somehow picked out by the mental concept employed in the explanation and the action explained? This assumption will seem obvious if you take a relational approach to causation. If all causation is relational, then explanations that reveal causal information will reveal information about causal relations, because what other kind of causal information is there?

Although both causalists and non-causalists assume a relational approach to causation, I think this assumption is more supportive of the Davidsonian/causalist position. This is because, although I agree with non-causalists that mental concepts like *belief* and *desire* do not seem to designate causally efficacious *items*, I think the intuition that rationalising explanations are causal is hard to resist. This means there is a strong motivation to accommodate valid points made by the non-causalists, without giving up the idea that rationalising explanations are causal.

Davidson's anomalous monism lets one do this. Davidson thinks that mental concepts are anomalous, which is to say that they are unsuitable for inclusion in causal laws of the form: 'there is an event-kind F, of which the cause event is a token, and an event-kind G, of which the effect event is a token, such that F events always cause G events'. This means that Davidson can acknowledge that there are significant differences between the explanatory scheme of explanations of actions that employ mental concepts and typical, scientific causal explanations. (The latter, Davidson thinks, do imply causal laws.) For instance,

Davidson can agree with non-causalists that mental concepts do not *seem* to perform their explanatory function by designating causes, because as *mental* concepts we should not expect them to. The anomalousness of mental concepts means that the causal nature of mental states and events is not revealed when these entities are picked out by mental concepts. This does not mean, however, that the facts that make that rationalising explanation genuinely explanatory are not causal facts. As Erasmus Mayr puts it:

> For Davidson, the epistemological criteria that we use for determining for which reason an agent has acted are the considerations of rationality and overall coherence among his mental states that are generally relevant for the interpretative enterprise of 'making sense of the agent'. What makes the reasons-explanation true, however, is something completely different: the obtaining of an event-causal link between reason and action, which for Davidson must be based on a strict causal law. (2011: 269–270)

The causalist can thus argue that, even though mental concepts do not *seem* to perform their explanatory function by designating causes, rationalising explanations would not be true if mental concepts did not somehow pick out events that stand in causal relations to actions. Of course, anomalous monism might not be correct, but I think that the opposition between Davidson and non-causalists on the matter of rationalising explanations is at something of an impasse, because anomalous monism is an available position.

The causal theory of action explanation and the causal theory of action are part of what is called 'the standard story' of human action. The theories are intuitively plausible enough to have become the standard account of what intentional action is and how it is explained, the account other theories must be weighed against. This is so despite the fact that causal theories of intentional action suffer some significant shortcomings, which I will discuss in Chapter 4. Why do causal theories of intentional action enjoy such intuitive plausibility? I contend that causal theories of intentional action seem superior to alternatives in part because philosophers of action assume a relational approach to causation. It is very difficult to imagine an alternative understanding of the causality of intentional action if you take a relational approach to causation. If causation is always, everywhere a relation between events, it would seem that the causation demonstrated in intentional action must be a relation, because all causation is, and will count as mental causation if and only if at least one of the terms of that relation is a mental entity. Furthermore, if causation is always, everywhere a relation between events, then explanations that reveal causal information will reveal information about causal relations. Thus, if rationalising explanations are causal, then they must point to causal relations between items somehow picked out by the mental concepts employed in the explanation and the action explained.

References[10]

Alvarez, M 2010 *Kinds of reasons: An essay in the philosophy of action*. Oxford: Oxford University Press.

Bishop, J 1989 *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Bokulich, A 2011 How scientific models can explain. *Synthese*, 180(1): 33–45. DOI: https://doi.org/10.1007/s11229-009-9565-1

Bratman, M 2001 Two problems about human agency. *Proceedings of the Aristotelian Society*, 101(3): 309–326. DOI: https://doi.org/10.1111/j.0066-7372.2003.00033.x

Brent, M 2017 Agent causation as a solution to the problem of action. *Canadian Journal of Philosophy*, 47(5): 656–673. DOI: https://doi.org/10.1080/00455091.2017.1285643

Dancy, J 2000 *Practical reality*. New York: Oxford University Press.

Davidson, D 1963 Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700. DOI: https://doi.org/10.2307/2023177. Reprinted in Davidson 2001 pp. 3–20.

Davidson, D 1971 Agency. In: Binkley, R, Bronaugh, R and Marras, A *Agent, action, and reason*. Toronto: University of Toronto Press. pp. 1–37. Reprinted in Davidson 2001 pp. 43–62.

Davidson, D 1973 Freedom to act. In: Honderich, T *Essays on freedom of action*. New York: Routledge and Kegan Paul. pp. 137–156. Reprinted in Davidson 2001 pp. 63–82.

Davidson, D 1987 Problems in the explanation of action. In Smart, J J C et al. *Metaphysics and morality: Essays in honour of J.J.C. Smart*. New York: Blackwell. pp. 35–49.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Enç, B 2003 *How we act: Causes, reasons, and intentions*. New York: Oxford University Press.

Frankfurt, H 1969 Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23): 829–839. DOI: https://doi.org/10.2307/2023833

Frankfurt, H 1978 The problem of action. *American Philosophical Quarterly*, 15(2): 157–162.

Hempel, C and Oppenheim, P 1948 Studies in the logic of explanation. *Philosophy of Science*, 15(2): 135–175. DOI: https://doi.org/10.1086/286983

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

[10] Author note: some references to Davidson are formatted (1963/2001). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001 collection of his essays, with the page numbers relating to that volume.

Kitcher, P 1989 Explanatory unification and the causal structure of the world. In Kitcher, P and Salmon, W *Scientific explanation*. Minneapolis, MN: University of Minnesota Press. pp. 410–505.

Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.

Mele, A 1992 *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.

Mele, A 2003 *Motivation and agency*. Oxford: Oxford University Press.

Raley, Y 2007 The facticity of explanation and its consequences. *International Studies in the Philosophy of Science*, 21(2): 123–135. DOI: https://doi.org/10.1080/02698590701498035

Searle, J 1983 *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.

Smith, M 2004 The structure of orthonomy. *Royal Institute of Philosophy Supplement*, 55: 165–193. DOI: https://doi.org/10.1017/s1358246100008675

Smith, M 2012 Four objections to the standard story of action (and four replies). *Philosophical Issues*, 22(1): 387–401. DOI: https://doi.org/10.1111/j.1533-6077.2012.00236.x

Smith, M 2021 Are actions bodily movements? *Philosophical Explorations*, 24(3): 394–407. DOI: https://doi.org/10.1080/13869795.2021.1957205

Tanney, J 2009 Reasons as non-causal, context-placing explanations. In: Sandis, C *New essays on the explanation of action*. Basingstoke: Palgrave Macmillan. pp. 94–111.

Van Fraassen, B C 1980 *The scientific image*. Oxford: Clarendon Press.

Velleman, D J 1992 What happens when someone acts? *Mind*, 101(403): 461–481. DOI: https://doi.org/10.7591/9781501721564-008

Woodward, J 2003 *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

CHAPTER 3

The Relational Approach to Causation

I turn now to the third aspect of the physicalist triad: the relational approach to causation. A theory of causation is relational if and only if it is committed to the following thesis:

> **Relationalism:** causation is always and everywhere a relation between distinct entities ('cause' and 'effect'); the worldly phenomenon that is referred to by our concept 'causation' is not ontologically diverse in this respect.

We have seen how the relational approach to causation lends plausibility to both physicalism and causal theories of intentional action. The driving force behind arguments for physicalism is the problem of mental causation, but the way mental causation is understood in these debates is heavily influenced by background assumptions about the nature of causation. Specifically, philosophers writing on the problem of mental causation assume that mental causation is a cause–effect relation where the cause relatum or effect relatum, or both, is a mental item (the relational understanding of mental causation). It is very difficult to imagine an alternative understanding of mental causation if you take a relational approach to causation. On this approach, 'cause' is an unequivocal term. All causation everywhere is the same, so the only thing that can discriminate between different categories of causation is the nature of the relata involved. The relational approach to causation also entails that causal reality is nothing more than a chain of causally related events, so, if intentional action is a causal phenomenon at all, it must be located within this worldview. This lends support to causal theories of intentional action that reduce the agent's role in bringing about what she intends to causation by mental events.

The relational approach to causation is not argued for by physicalists or those who propose a causal theory of intentional action. Instead, it is often taken for granted or treated as a harmless background assumption or

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 55–71. York: White Rose University Press. DOI: https://doi.org/10.22599/White.d.

pre-theoretical notion. I think this is incorrect, because I think the assumption that causation is always and everywhere a relation is not as innocuous as it seems. It is a substantive claim about the nature of causation. The purpose of this chapter is to show that the relational approach to causation is a substantive philosophical position, and not merely a harmless background assumption or pre-theoretical notion.

3.1 Hume's legacy

The relational approach to causation is not recognised as a substantive philosophical position because most philosophers working on causation accept relationalism, at least implicitly. Most have assumed that providing a theory of causation is a matter of explaining what a relation must be like to be a causal relation. In his Stanford Encyclopedia article on the metaphysics of causation, Jonathan Schaffer introduces this philosophical project with the following question: 'What must a world be like, to host causal relations?' (2016). He goes on to state that '[q]uestions about the metaphysics of causation may be usefully divided into questions about the causal relata, and questions about the causal relation' (2016). The majority of work on the metaphysics of causation proceeds as if Schaffer's taxonomy of questions concerning causation are the only questions we can ask about what reality must be like when causal statements are true. In J. Dmitri Gallow's Stanford Encyclopedia article on the metaphysics of causation, which replaced Schaffer's article, the metaphysics of causation is still described as the project of finding out 'what kind of relation [causal] claims are about' (2022). The possibility that causation may not fit into a single ontological category is rarely taken seriously.[11]

Relationalism is widely accepted in part due to the lasting influence David Hume has had on the philosophy of causation. Briefly examining Hume's influence on the philosophy of causation will help make it clear that, far from being pre-theoretical, relationalism has its roots in Humean theories of causation.

During the early modern period, the concept 'cause' underwent a transformation. Earlier Aristotelian and Scholastic ideas about causation were challenged, replaced, and abandoned, including the Aristotelian view that there are four modes of causation, or that 'cause' has four distinct senses. Hume concluded that 'all causes are of the same kind, and that in particular there is no foundation for that distinction, which we sometimes make betwixt efficient causes and causes *sine qua non*; or betwixt efficient causes, and formal, and material, and exemplary, and final causes' (1964: 171). Since Hume, philosophers of causation have come to regard efficient causation as the only mode of

[11] One notable exception is Helen Steward (2012: 212–216), whose consideration of the ontological heterogeneity of causal reality informs her understanding of agency.

causation there is: all causation is a matter of causes being that which produced effects. Aristotle's other modes of causation are not really causation at all; they are more accurately described as modes of explanation or modes of 'because' (Hocutt 1974). Following Hume, contemporary philosophy of causation rarely entertains the idea that 'cause' might be ambiguous. The univocality of the concept 'cause' is a key tenet of relationalism. Relationalism entails that, when we inquire about what reality must be like when true causal statements are made, there is just one sort of thing we are looking for—it is not the case that the reality causal statements answer to might vary depending on the context within which those statements are made.

Aristotelian ideas about substances and powers and how these concepts figure in causation were also challenged during the early modern period. As Walter Ott (2009) describes it, Aristotelian ideas about substances and powers were gradually replaced by laws of nature and a mechanist ontology (albeit in a messy, often piecemeal way), a development that abetted Hume's scepticism about the existence of a mind-independent necessary connection between cause and effect.

Hume wanted to know what the source or origin of our idea of necessary connection was and argued forcefully that we gain no impression of it when we observe a single instance of one type of event being followed by another. Hume argued that we experience 'one event follows another; but we never can observe any tie between them. They seem conjoined, but never connected' (1975: 74). Hume drew a similar conclusion with regard to powers: we observe 'an uninterrupted succession' but not any 'power or force which actuates the whole machine' (1975: 63); we can perceive what a thing is like but not what it is capable of doing. Hume argued further that we cannot perceive the operation of power even in cases where we ourselves are doing something or making something. Even in these cases, all we observe is a sequence of events. However, Hume argued, when we *repeatedly* experience events of one type being followed by events of another type, we come to *expect* an event of the second type when we experience an event of the first, and this internal feeling of expectation is the impression from which this idea of necessitation between cause and effect arises. On one interpretation, Hume's conclusion is that the idea of causation as necessary connection or the exercise of power is a product of our own minds, and what exists in mind-independent reality are unconnected events within which we can discern patterns of regularity.

This admittedly controversial interpretation of Hume has had a lasting influence over modern theories of causation.[12] The principle that cause and effect are distinct events and so there can be no metaphysically necessary connections between them, a principle sometimes known as 'Hume's dictum' (Wilson 2010), presents a challenge. If cause and effect are not joined by a necessitating

[12] See Beebee (2007) and Millican (2007) for good discussions on how Hume's claims about causation should be interpreted.

relation, how are they joined? *This* is the challenge contemporary theories of causation have focused on. As a result, the project of giving an account of the metaphysics of causation has become a matter of specifying the nature of the relation that joins cause and effect together. Relationalism is taken for granted by many contemporary theories of causation, and the ontologically richer views of causation entertained by Aristotelians and Scholastics rarely surface in modern theories of causation. However, the fact that Aristotelian ideas, such as the idea that there are different kinds of cause, stand opposed to relationalism shows that relationalism is not a pre-theoretical assumption about the nature of causation. Relationalism is the dominant theoretical position within contemporary philosophy of causation, but it is still a theoretical position.

3.2 Two relational theories of causation

The regularity theory of causation and David Lewis's counterfactual theory of causation are paradigm examples of relational theories of causation. Both theories hold that causation is a special type of relation between cause and effect. Both theories also attempt to spell out what this special relation is in non-causal terms. In that way, both theories offer a reductive account of causation. Briefly examining the metaphysical commitments of these theories will help make it clear what beliefs about causation are consistent with relationalism. It will also make it easier to articulate the alternative to relationalism in later chapters.

The regularity theory holds that causation, as it exists in the world independently of our thinking about it or knowledge of it, is exhaustively constituted by certain relations of spatiotemporal contiguity that obtain with regularity. More specifically, the regularity theory holds that causation is a relation of spatiotemporal contiguity between two events, *c* and *e*, where *c* occurs before *e*, and where all events of the same type as *c* are regularly followed by events of the same type as *e*. The regularity theory as stated above faces problems and, in response, more sophisticated versions of the theory have been proposed.[13] However, the simplest version of the regularity theory will suffice for my purposes here.

The main argument for adopting a regularity theory is that it offers a reductive account of causation where, as Stathis Psillos puts it, 'causal talk becomes legitimate, but it does not imply the existence of a special realm of causal facts that make causal talk true, since its truth conditions are specified in non-causal terms, that is, in terms of spatiotemporal relations and actual regularities' (2002: 4). The idea is that the regularity theory of causation—or at least a suitably worked-up version of it—provides everything we would want from a theory of causation, without positing the existence of powers or a *sui generis* kind of

[13] For example, Baumgartner (2008), Mackie (1974) and Mill (1843) have all offered more sophisticated versions of the regularity theory.

necessity. According to the regularity theory, what ascriptions of power, or statements about what a thing can do, actually *mean* (if they are not false or nonsense) is that the behaviour of the object to which the 'power' is attributed is regular in a certain way. That is, it might be true to say some object has a power, but what makes such a statement true will be some fact about the arrangement of the spatiotemporal mosaic of instantiations of intrinsic, qualitative, categorical properties.

The mosaic metaphor is how Lewis describes the metaphysics presupposed by the regularity theory. In more detail, this metaphysics says:

> [I]n a world like ours, the fundamental relations are exactly the spatiotemporal relations: distance relations, both spacelike and timelike, and perhaps also occupancy relations between point-sized things and spacetime points. And it says that in a world like ours, the fundamental properties are local qualities: perfectly natural intrinsic properties of points, or of point-sized occupants of points. Therefore it says that all else supervenes on the spatiotemporal arrangement of local qualities throughout all of history, past and present and future. (1994: 474)

As Lewis puts it in the introduction to his *Philosophical Papers (vol. II)*, 'all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another' (1986: ix). Jonathan Schaffer describes this worldview slightly differently: Schaffer writes that the world is 'history' i.e. 'the fusion of all events throughout space-time' (2007: 83).

Lewis's (1973a; 1973b) counterfactual theory of causation analyses causation in terms of counterfactual dependence. This theory exploits the intuition that causes are that which made the difference to the occurrence of the effect; that is, had the cause not occurred, the effect would not have occurred either. Lewis developed this idea by analysing the causal relation as the ancestral of a counterfactual dependence relation. So, an event *c* stands in a causal relation to another event *e* if and only if *e* counterfactually depends on *c*, or *e* counterfactually depends on an event that counterfactually depends on *c*, or *e* counterfactually depends on an event that counterfactually depends on an event that counterfactually depends on *c*, etc. As with the regularity theory, Lewis's counterfactual theory has been modified in light of objections raised against the original version, but again the simplest version of the counterfactual account will suffice for now.[14]

Lewis's counterfactual theory's status as reductive depends, in part, on Lewis's theory of modality. Lewis opts for a possible world semantics for counterfactuals. So, a counterfactual like 'if *c* had not occurred, then *e* would not have occurred' is true if and only if *e* does not occur at the closest possible

[14] See, for example, Ganeri, Noordhof and Ramachandran (1996), Lewis (2000), McDermott (2002) and Sartorio (2005).

world where *c* does not occur. How close a possible world is to the actual world depends on how *similar* that world is to the actual world. For Lewis, similarity between two possible worlds is determined by what particular states of affairs obtain at the two worlds and what the laws of two worlds are. So, world w^1 is more similar to world w^2 the more states of affairs w^1 has in common with w^2 and the more laws w^1 has in common with w^2.

If one went along thus far with Lewis's semantics for counterfactuals, but thought that laws of nature were brute facts about what powers things have, or facts about primitive 'necessitation' relations holding between universals, then even if one opted for an account of causation where causation is reduced to counterfactual dependence, the resultant theory of causation would not be reductive. This is because, on such a view, the truth of counterfactual conditionals depends on similarity rankings of possible worlds, which in turn depends on brute facts about powers, or a *sui generis* form of necessity. However, Lewis gives an account of laws of nature that does not presuppose the existence of powers or anything over and above the spatiotemporal mosaic of instantiations of intrinsic, qualitative, categorical properties. For Lewis, laws of nature are simply regularities that are deducible from axioms in an explanatory system that best balances simplicity and strength. An explanatory system picks as few general truths as possible to serve as axioms—the fewer, the simpler—then deductively derives further general truths from these. The more general truths the system deductively entails, the stronger the system. As Helen Beebee (2006) points out, because Lewis seeks to analyse causation without assuming the existence of any kind of worldly necessitation, and ends up turning to regularities in order to fulfil that mandate, Lewis's counterfactual theory of causation has a lot in common, metaphysically speaking, with the regularity theory. On both theories, the worldly structures that make true causal claims are, in the end, regularities. And, just like the regularity theory, Lewis's counterfactual theory does not posit any kind of entity or deeper fact (like facts about what powers things have or what is a natural necessity) that grounds or explains why regularities hold, or why certain counterfactual conditionals are true.

What is important to notice about these two theories is that they reject the idea that causation is (at least sometimes) the exercise of power or the making-happen of an effect in favour of describing causation in more ontologically sanitised terms, which they assume means describing causation in terms of a relation between elements of the 'spatiotemporal mosaic'. Many rivals to the regularity theory or the counterfactual theory of causation challenge the *reductive* aspects of these theories. That is, rival theories of causation challenge the principle that causation, as it exists in the world independently of our thinking about it or knowledge of it, is exhaustively constituted by non-causal states of affairs. However, the principle that causation is always a relation between cause and effect is not challenged.

For example, Galen Strawson articulates a conception of causation that he calls Causation with a capital 'C'. To believe in the existence of Causation is to believe: 'a) that there is something about the fundamental nature of the world in virtue of which the world is regular in its behaviour; and b) that that something is what causation is, or rather it is at least an essential part of what causation is' (1989: 84–85). Strawson thus advocates a view that takes causation to be an entity that grounds the world's regularities but cannot be reduced to regularities, or indeed any aspect of the 'spatiotemporal mosaic'. Although Strawson (1989) argues that causation, as it is in reality, is regular succession plus something extra, which explains why events unfold in a regular way, he is noncommittal on what this extra element is. Strawson (1987) suggests that this additional element could be the presence of 'objective forces—e.g. the "fundamental forces" postulated by physics' that 'govern the way objects behave and interact' (1987: 254), and adds:

> I will avoid speaking of 'natural necessity', or of 'laws of nature' (understood in a strong, non-Regularity-theory sense), or of the 'causal powers' of objects. It is very difficult to keep control of these rival terminologies. But here the notion of objective forces is being understood in such a way that accounts of causation given in terms of these other notions may be supposed to reduce naturally to the account in terms of forces. For example: (1) if objects have causal powers, they have the powers they do wholly in virtue of the nature of the forces informing (and so governing) the matter of which they are constituted. (1987: 255)

Michael Tooley (1990a) similarly argues against views that hold that 'causal relations are … logically supervenient upon non-causal properties and relations' (1990a: 217). The sort of causal realism that Tooley endorses treats 'causal concepts as theoretical concepts, so that causal relations can only be characterised, indirectly, as those relations that satisfy some appropriate theory' (1990a: 234). The appropriate theory, Tooley (1990b) proposes, is one that includes claims about the formal properties of causal relations, and which tells us what a law must be like to be a causal law. Causal relations are thus relations that have the right formal properties and 'whose presence in a law makes that law a causal one' (1990b: 303). Tooley shares Armstrong's view about laws of nature (of which causal laws are a subset); that is, he thinks that laws are necessitation relations between universals. So, it would seem that Tooley's account of causation, in virtue of its appeal to causal laws, makes use of a *sui generis* form of necessity.

The point I wish to emphasise is that both Strawson and Tooley are arguing specifically against attempts to reduce the causal *relation* to some non-causal *relation*. Strawson (1989) is concerned with showing that we should believe there is something more to the relation between cause and effect than regular

succession. Similarly, it is specifically 'realism with regard to causal relations' that Tooley considers (1990a: 233).

Interestingly, Tooley cites Elizabeth Anscombe as a philosopher who upholds a realist view of causation where causal relations are directly observable 'not only in the everyday sense of that term, but in a much stronger sense which entails that concepts of causal relations are analytically basic' (1990a: 233–234). Anscombe (1971) suggested that we come by our primary knowledge of causality when we learn to speak and come to associate the linguistic representation of a causal concept with its correct application. An example of such a causal concept that Anscombe provides is 'infect'. Others include 'scrape, push, wet, carry, eat, burn …' (1971: 9). She suggests that causal activities like scraping and pushing (though perhaps not infecting) are activities that we can directly perceive. Tooley ultimately rejects this form of realism. He argues that, even if Anscombe is right that we know by observation that one thing is pushing another (for example), this does not show that what it is about the events we are seeing that means they are causally related is something irreducible we can nevertheless observe. It might be that we infer, from what we perceive, that causation is there.

However, I think that Tooley has misconstrued what Anscombe is claiming in her 1971 lecture 'Causation and Determination', from which he cites. What Anscombe suggests we directly perceive is not a special *relation* between cause and effect but substances exerting causal power over other substances. We do not observe a cause causing an effect; we observe an agent acting on a patient. Anscombe is suggesting that an agent acting on a patient is causation, and this is in spite of the obvious truth that agent and patient are not related to each other as cause and effect. Anscombe's point is that we come by knowledge of causality when we directly perceive agents pushing patients and correctly associate what we see with the inherently causal concept 'pushing'. Tooley might be right that the fact that we directly perceive agents pushing patients (for example) may not be enough to show that we directly perceive a connection between the events that makes it the case that they are causally related. But why can't the fact that we directly perceive an interaction be enough to show that we directly perceive causation? Tooley construes Anscombe's claim incorrectly, I think, because of his commitment to a version of relationalism that says that causation is a relation between events.

3.3 Manipulability accounts of causation

Another important family of theories of causation is manipulability accounts of causation. Manipulability accounts of causation explore the intuition that causes are things in nature that we can manipulate and thereby alter outcomes. These theories connect causation to our sense of agency, to the idea of ourselves as beings which alter the course of events. Indeed, some manipulability

accounts explicitly define causation in terms of agency. For example, Georg Henrik von Wright argues that an event *c* is the cause of event *e* if and only if bringing about *c* is a way for an agent to bring about *e*, that is, only if *e* can be considered the result of the action of bringing about *c*:

> [T]o think of a relation between events as causal is to think of it under the aspect of (possible) action. It is therefore true, but at the same time a little misleading to say that if *p* is a (sufficient) cause of *q*, then if I could produce *p* I could bring about *q*. For *that p* is the cause of *q*, I have endeavoured to say here, *means* that I could bring about *q*, if I could do (so that) *p*. (1971: 74)

Similarly, Peter Menzies and Huw Price argue that 'an event [*c*] is cause of distinct event [*e*] just in case bringing about the occurrence of [*c*] would be an effective means by which a free agent could bring about the occurrence of [*e*]' (1993: 187) and an event *c* is an effective means by which a free agent could bring about occurrence of *e*, just in case the probability of *e* occurring given that *c* was brought about by a free agent is greater than the unconditional probability of *e* occurring.

In assigning a central role to human agency, these theories might seem to offer a richer account of causation, one that leaves room for the idea that causation could be something other than a relation between cause and effect; instead, it might be an activity (manipulation) that agents perform, or it might be the exercise of power where this is an irreducible feature of fundamental reality. However, closer examination of manipulability theories reveals that most are committed to relationalism.

A criticism levied against agency-based manipulability accounts is that they are problematically circular, because agency is a causal notion: *producing* and *bringing about* are causal concepts, hence agency-based theories purport to analyse causation in terms of causation. Von Wright responds to this objection by arguing that the relation between an action (e.g. cutting of the cake) and its result (the cake's coming to be cut) is *not* a causal relation; it is rather a logical one (if the cake does not come to be cut, then no-one cut it—the cutting-of-the-cake action did not take place):

> I am anxious to separate agency from causation. Causal relations exist between natural events, not between agents and events. When by doing *p* we bring about *q*, it is the happening of *p* which causes *q* to come. And *p* has this effect quite independently of whether it happens as a result of action or not. (1974: 49)

I think von Wright is right to sharply distinguish between agency on the one hand and causal relations on the other—he is correct that to demonstrate agency is not for an agent to stand in a causal relation to an event. However, I

do not think, as von Wright does, that this entails that agency is not a causal phenomenon. Von Wright does not recognise this because he subscribes to relationalism, the view that causation is always, everywhere a relation. Von Wright's view can be thought of as abiding by the following reasoning: causation is the relation between cause and effect; agency is not a relation between cause and effect; therefore, agency is not causation. This argument is sound only if relationalism is true. So, von Wright accepts relationalism.

The circularity objection can be directed against Menzies and Price's view as well. Menzies and Price respond to the circularity objection in the following way:

> The basic premise is that from an early age, we all have direct experience of acting as agents. That is, we have direct experience not merely of the Humean succession of events in the external world, but of a very special class of such successions: those in which the earlier event is an action of our own, performed in circumstances in which we both desire the later event, and believe that it is more probable given the act in question than it would be otherwise. To put it more simply, we all have direct personal experience of doing one thing and thence achieving another … It is this common and commonplace experience that licences what amounts to an ostensive definition of the notion of 'bringing about'. In other words, these cases provide direct non-linguistic acquaintance with the concept of bringing about an event; acquaintance which does not depend on prior acquisition of any causal notion. An agency theory thus escapes the threat of circularity. (1993: 194–195)

Unlike von Wright, Menzies and Price do not deny that agency is a causal phenomenon. What they deny is that acquiring the agency concept requires that one has already acquired the concept of causation. For Menzies and Price, even though agency itself is an essentially causal phenomenon, the *concept* of agency is one that can be understood and grasped independently of the *concept* of causation, and, because it can be independently understood, it can be used to analyse causation. As for whether Menzies and Price accept relationalism, it is not exactly clear. They describe the agency concept as 'a special class of successions' and as an action causing a result, which seems to suggest that they view agency in relational terms. However, ultimately I think it is unclear whether Menzies and Price's version of a manipulability account of causation accepts relationalism or not.

James Woodward (2003) argues that Menzies and Price's view is unacceptably anthropomorphic and subjectivist. Because Menzies and Price invoke a concept of agency that we grasp via direct experience of our own agency at work, their theory faces a difficult problem concerning causes that cannot be manipulated by human agents. To take an example from Menzies and Price (1993: 195), it seems to be true that movement of tectonic plates caused the 1989 San Francisco earthquake, but it is not true that movement of tectonic

plates was an event that could have been an effective means by which a human agent could have brought about the earthquake. Manipulating tectonic plates is just not within our power.

Woodward (2003), building on work by Judea Pearl (2000), offers his own manipulability theory of causation, which avoids this problem by using the concept of an intervention to analyse the causal relation, rather than manipulation by a human agent. Woodward contends that a variable *c* is causally related to a variable *e* if and only if intervention on *c* leaves the relationship between *c* and *e* invariant but changes the value of *e*. An intervention is any event that 'surgically' causes the value of *c* to change, that is, by blocking all causal influence over the value of *c* the usual causal antecedents of *c* have and without causally influencing the value of *e* except through *c*. An intervention is any event that has certain causal characteristics; an intervention need not involve human agency at all (although no doubt many interventions do involve human agency).

Woodward's theory is a kind of counterfactual theory of causation, since whether two variables are causally related to each other depends on how the relationship between those variables would change if certain interventions were made. However, there are key differences between Woodward and Lewis when it comes to the semantics of counterfactual conditionals. The most important difference is that in Lewis's account of how we should evaluate counterfactual conditionals in causal contexts it is never necessary to appeal to causal facts. By contrast, in Woodward's account of how we should evaluate counterfactual conditionals in causal contexts we are supposed to imagine that the antecedent of the counterfactual is made true by the occurrence of an intervention, which presupposes that certain causal facts obtain. To illustrate this point with an example, suppose event *c* caused e_1 and e_2, and e_1 and e_2 are not causally related to each other. Because counterfactual dependence is sufficient for causation, we would want the following counterfactual to come out false:

(a) If e_1 had not occurred, e_2 would not have occurred.

But, in a world where e_1 does not occur, we might suppose that this was because it was not caused by *c*, i.e. because *c* did not occur—but in that case e_2 would not have occurred either. This world—where e_1 does not occur because *c* does not occur—is therefore the wrong world to turn to when evaluating the truth of the counterfactual in a causal context. Lewis recommends that when we evaluate counterfactuals in a causal context we forbid 'backtracking'—i.e. we are forbidden from imagining that prior events and circumstances were also changed so as to cause the antecedent of our target counterfactual to be true. When we evaluate (a) we must imagine that a small miracle makes it the case that e_1 does not occur. So, the world we should use to evaluate the truth of (a) is a world where *c* still happens but then, miraculously, e_1 does not occur—in such a world e_2 would still occur (because *c* would still cause it), and therefore (a) comes out false.

Woodward achieves this same result using the notion of an intervention, rather than the notion of a 'small miracle'. For Woodward, when we evaluate (a) we are supposed to imagine that an intervention occurred to make it the case that e_1 did not occur—and such an intervention, by definition, leaves all causal relationships, except those which have e_1 as effect, unchanged. Evaluating the truth of (a) thus requires assuming certain other causal relations in the situation under discussion obtain. Even though Woodward's and Lewis's theories differ in this important way, it is not part of Woodward's theory that the truth of counterfactual conditionals depends on brute facts about powers, or a *sui generis* form of necessity. Thus, Woodward's theory is consistent with the view that counterfactual dependence can be understood without a primitive concept of power.

Does Woodward's theory embrace relationalism? The theory is intended to identify causal relationships between variables. On this theory, causation is something that exists between nodes in a network, and the concept of an intervention can tell us which relationships within this network are genuinely causal. On Woodward's theory, there is nothing extra in addition to the relationships between variables—such as the exercise of causal power or the bringing-about of events—which is essential to our understanding of causation. For this reason, I consider Woodward's theory a relationalist theory.

3.4 The relata of causation

Relationalism says that causation is always and everywhere a relation between distinct entities; however, it does not prescribe anything specific about what these entities must be. There is great disagreement on what the relata of causation are. There are many who hold that causation is a relation between *events* (Davidson 1967; Kim 1976; Lewis 1986). Some philosophers think that the relata of causation are *facts* (Bennett 1988; Mellor 1995). As mentioned above, Woodward (2003) holds that causation holds between *variables*. It has also been suggested that causation holds between *states of affairs* (Armstrong 1997), *conditions* (Mackie 1965) and *tropes* (Ehring 2011). I doubt this list is exhaustive. The situation is further complicated by the fact that there is very little agreement on the nature of entities like events, facts and states of affairs.

For example, among those who agree that causation is a relation between events, there is disagreement on what exactly events are. Davidson thinks that events are concrete particulars that can be redescribed and reidentified under different modes of presentation. This means that one and the same event can be referred to via different expressions, each of which identifies the event via a different intrinsic feature of it. For example, on Davidson's conception of events, Boudicca's death, Boudicca's suicide and Boudicca's poisoning are all one and the same event identified with different descriptions.

By contrast, Kim (1976) takes events to be 'exemplifications of properties at times'. Kim-events are located in space (they are where the objects exemplifying the properties are), and they are bound to a particular time (the times at which, or during which, the object exemplifies the properties) and they are contingent (they exist only if some object is a certain way). For Kim, the fact that his property exemplifications are bound to a particular time means that his events are particulars. However, Kim-events are also fact-like. Like facts, Kim-events indicate that an object is qualified. Also like facts, Kim-events have a propositional structure. The structure of Kim-events means that Kim-events are much more fine-grained than Davidsonian events. For example, Boudicca's exemplifying the property dying by suicide and Boudicca's exemplifying the property dying by poisoning would be distinct events as they involve distinct properties (dying by suicide and dying by poisoning).

Although relationalism is technically neutral with regard to what the relata of causation are, relational theories of causation that take causation to be a natural, extensional relation that holds between particulars (even fine-grained, fact-like particulars) are more supportive of other elements of the physicalist triad than theories of causation that allow causation to be an intentional relation. To see this, recall that the argumentative force of the causal argument is that, if mental and physical items are distinct, then they are in competition with each other for status as the cause of a physical effect. In order for there to be competition here, whether the mental item is cause of the physical effect cannot be something that depends on how the physical effect is described. The causal connection between the physical effect and its cause has to be a real relation. Furthermore, proponents of causal theories of intentional action state that their aim is to naturalise agency. As Bishop states, causal theories of action promise to 'make intelligible the possibility of agency within the natural order' (1989: 10). And Enç describes the causal theory of action as a 'treatment of action that confines itself just to events of the natural order of things, and to the causal relations among them' (2003: 3). Explaining agency in terms of events and causal relations could only be considered a project of naturalisation if causal relations are themselves natural relations that exist 'out there in the world'. As Giuseppina D'Oro explains:

> It is only if the term 'causation' is taken to be a category of revisionary metaphysics denoting a real relation, holding amongst events independently of how they are described, that the problem of causal rivalry between folk-psychological explanations of actions and naturalistic explanation of events can arise. The problem of explanatory exclusion simply does not arise within a descriptive conception of metaphysics precisely because, within such a conception of the role and character of philosophical analysis, causal relations are intentional relations that are not logically independent of the explanatory goals of a science. (2012: 219)

3.5 The importance of relationism

A central claim of this book is that the relational approach to causation is one of three mutually supporting views that form the physicalist triad. The relational approach to causation is, in some ways, the most fundamental of these three elements. The relational approach is appealing to both physicalists and those who endorse a causal theory of intentional action because of its associations with naturalism. The relational approach also lends support to both physicalism and causal theories of intentional action. If one adopts a relational approach to causation, then it seems inevitable that mental causation will be understood in relational terms, i.e. presented as a cause–effect relation between mental and physical entities. If all causation everywhere is the same, the only thing that can discriminate between different categories of causation is the nature of the relata involved. Furthermore, if one adopts a relational approach to causation, intentional action must be distinguished in terms of its aetiology. Alternatives to causal theories of intentional action, which purport to understand intentional action in terms of irreducible agent causation, are uncongenial to the relational approach to causation.

As I have already mentioned, I am not the first to suggest that there are intellectual connections between physicalism, philosophy of action and philosophy of causation (see for example Hornsby 2015; Lowe 2008). However, as Jennifer Hornsby (2015) notes, these connections have been underexplored. Some writers in philosophy of mind have suggested that the best way to respond to the causal argument for physicalism is to challenge the assumptions about causation implicit in the argument. For example, List and Menzies (2009) argue that construing causation as 'difference-making' allows one to argue that higher-level mental properties are not causally excluded by the physical properties that realise them. However, what these writers suggest is a fairly modest rethinking of the assumptions about causation at work in the causal argument, and the metaphysics of mind they eventually endorse is usually a kind of non-reductive physicalism. A number of writers in philosophy of action who are dissatisfied with causal theories of intentional action have suggested that Aristotelian views about causation are needed to properly understand agency. However, as we shall see in Chapter 5, although these neo-Aristotelian views of agency posit the existence of a special kind of causation (agent causation or substance causation), they do not explicitly challenge the idea that causation is always, everywhere a relation.

In the last three chapters I have tried to make salient the mutually supporting relationships between physicalism, causal theories of intentional action and the relational approach to causation. My next task is to explain why the best strategy for resisting the conclusion of the causal argument for physicalism is to use lessons from philosophy of action to challenge the relational understanding of mental causation.

References

Anscombe, G E M 1971 *Causality and determination: An inaugural lecture*. Cambridge: Cambridge University Press.

Armstrong, D M 1997 *A world of states of affairs*. Cambridge: Cambridge University Press.

Baumgartner, M 2008 Regularity theories reassessed. *Philosophia*, 36(3): 327–354. DOI: https://doi.org/10.1007/s11406-007-9114-4

Beebee, H 2006 Does anything hold the universe together? *Synthese*, 149(3): 509–533. DOI: https://doi.org/10.1007/s11229-005-0576-2

Beebee, H 2007 The two definitions and the doctrine of necessity. *Proceedings of the Aristotelian Society*, 107(1pt3): 413–431. DOI: https://doi.org/10.1111/j.1467-9264.2007.00231.x

Bennett, J 1988 *Events and their names*. Indianapolis, IN: Hackett.

Bishop, J 1989 *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

D'Oro, G 2012 Reasons and causes: The philosophical battle and the metaphilosophical war. *Australasian Journal of Philosophy*, 90(2): 207–221. DOI: https://doi.org/10.1080/00048402.2011.583930

Davidson, D 1967 Causal relations. *Journal of Philosophy*, 64(21): 691–703. Reprinted in Davidson 2001 pp. 149–162.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Ehring, D 2011 *Tropes: Properties, objects, and mental causation*. Oxford: Oxford University Press.

Enç, B 2003 *How we act: Causes, reasons, and intentions*. New York: Oxford University Press.

Gallow, J 2022 The metaphysics of causation. The Stanford Encyclopedia of Philosophy, 14 April 2022. Available at https://plato.stanford.edu/archives/sum2022/entries/causation-metaphysics [Last accessed 2 September 2023].

Ganeri, J, Noordhof, P and Ramachandran, M 1996 Counterfactuals and preemptive causation. *Analysis*, 56(4): 219–225. DOI: https://doi.org/10.1093/analys/56.4.219

Hocutt, M 1974 Aristotle's four becauses. *Philosophy*, 49(190): 385–399. DOI: https://doi.org/10.1017/s0031819100063324

Hornsby, J 2015 Causality and 'the mental'. *HUMANA.MENTE Journal of Philosophical Studies*, 8(29): 125–140.

Hume, D 1964 *A treatise of human nature. Volume 1*. London: J M Dent/E P Dutten.

Hume, D 1975 *Enquiries concerning human understanding and concerning the principles of morals*. 3rd ed. Oxford: Clarendon Press.

Kim, J 1976 Events as property exemplifications. In: Brand, M and Walton, D *Action theory*. Dordrecht: D. Reidel. pp. 310–326.

Lewis, D K 1973a Causation. *The Journal of Philosophy*, 70(17): 556–567. DOI: https://doi.org/10.2307/2025310
Lewis, D K 1973b *Counterfactuals*. Oxford: Basil Blackwell.
Lewis, D K 1986 *Philosophical papers*. New York: Oxford University Press.
Lewis, D K 1994 Humean supervenience debugged. *Mind*, 103(412): 473–490. DOI: https://doi.org/10.1093/mind/103.412.473
Lewis, D K 2000 Causation as influence. *Journal of Philosophy*, 97(4): 182–197.
List, C and Menzies, P 2009 Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy*, 106(9): 475–502. DOI: https://doi.org/10.5840/jphil2009106936
Lowe, E J 2008 *Personal agency: The metaphysics of mind and action*. New York: Oxford University Press.
Mackie, J L 1965 Causes and conditions. *American Philosophical Quarterly*, 2(4): 245–264.
Mackie, J L 1974 *The cement of the universe*. Oxford: Clarendon Press.
McDermott, M 2002 Causation: Influence versus sufficiency. *The Journal of Philosophy*, 99(2): 84–101. DOI: https://doi.org/10.5840/JPHIL200299219
Mellor, D H 1995 *The facts of causation*. London: Routledge.
Menzies, P and Price, H 1993 Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44(2): 187–203. DOI: https://doi.org/10.1093/bjps/44.2.187
Mill, J S 1843 *A system of logic, ratiocinative and inductive being a connected view of the principles of evidence and the methods of scientific investigation*. London: Longmans, Green, Reader, and Dyer.
Millican, P 2007 Against the new Hume. In: Read, R and Richman, K *The new Hume debate: Revised edition*. London: Routledge. pp. 211–252.
Ott, W 2009 *Causation and laws of nature in early modern philosophy*. Oxford: Oxford University Press.
Pearl, J 2000 *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.
Psillos, S 2002 *Causation and explanation*. London: Routledge.
Sartorio, C 2005 Causes as difference-makers. *Philosophical Studies*, 123(1/2): 71–96. DOI: https://doi.org/10.1007/s11098-004-5217-y
Schaffer, J 2007 Causation and laws of nature: Reductionism. In: Sider, T, Hawthorn, J and Zimmerman, D W *Contemporary debates in metaphysics*. Malden: Blackwell, pp. 82–107.
Schaffer, J 2016 The metaphysics of causation. The Stanford Encyclopedia of Philosophy, 5 July 2016. Available at https://plato.stanford.edu/archives/spr2022/entries/causation-metaphysics [Last accessed 2 September 2023].
Steward, H 2012 *A metaphysics for freedom*. Oxford: Oxford University Press.
Strawson, G 1987 Realism and causation. *The Philosophical Quarterly*, 37(148): 253–277. DOI: https://doi.org/10.2307/2220397
Strawson, G 1989 *The secret connexion: Causation, realism, and David Hume*. Oxford: Oxford University Press.

Tooley, M 1990a Causation: Reductionism versus realism. *Philosophy and Phenomenological Research*, 50: 215–236. DOI: https://doi.org/10.2307/2108040

Tooley, M 1990b The nature of causation: A singularist account. *Canadian Journal of Philosophy*, 20(1): 271–322. DOI: https://doi.org/10.1080/00455091.1990.10717229

von Wright, G H 1971 *Explanation and understanding*. Ithaca, NY: Cornell University Press.

von Wright, G H 1974 *Causality and determinism*. New York: Columbia University Press.

Wilson, J 2010 What is Hume's dictum, and why believe it? *Philosophy and Phenomenological Research*, 80(3): 595–637. DOI: https://doi.org/10.1111/j.1933-1592.2010.00342.x

Woodward, J 2003 *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

CHAPTER 4

Breaking Out of the Physicalist Triad

In the previous three chapters I outlined three philosophical positions that I believe are mutually reinforcing: (1) physicalism, (2) causal theories of intentional action and (3) relational approaches to causation. I have called this triad of views *the physicalist triad* because the consequence of endorsing each element of the triad is that physicalism about mentality becomes the only acceptable metaphysics of mind: it appears to be the only view that has a chance of saving the phenomenon of mental causation. In many arguments for physicalism, mental causation is understood in relational terms; that is, mental causation is presented as a cause–effect relation between mental and physical items. Philosophers writing about the problem of mental causation are limited to this way of describing what mental causation is because they assume that 'cause' is an unequivocal term—all causation everywhere is the same—so the only thing that can discriminate between different categories of causation is the nature of the relata involved. This assumption about causation, an assumption I have called 'relationalism', is ubiquitous in philosophy of causation but it is also a thesis that will be appealing to physicalists because of its associations with naturalism. Mental causation is also made to seem indispensable because of causal theories of intentional action. Causal theories of intentional action, however, owe their dominance to relational assumptions about causation. This is because, if causation is always, everywhere a relation, then explaining what intentional action is is a matter of distinguishing between different types of event causation (those that do and those that do not constitute intentional action). In summary, then, even though each element of the triad is logically independent, in practice they reinforce each other. Physicalists endorse relational approaches to causation because they are naturalistic; against the backdrop of the relational approach to causation, causal theories of intentional action are made to seem intuitively more appealing than their rivals; and endorsing causal theories of intentional action strengthens the case for physicalism by making relationally understood mental causation seem indispensable.

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 73–96. York: White Rose University Press. DOI: https://doi.org/10.22599/White.e.

In this chapter, I want to explain why I think we should try to break out of the physicalist triad. Physicalism, causal theories of intentional action and relational approaches to causation are individually plausible, with a lot of explanatory power. Physicalism purports to offer a naturalistic account of the mind; causal theories of intentional action promise to explain how it is that reasons can explain actions as well as offering an account of what makes the difference between what an agent does and what happens to him; relational approaches to causation, such as the regularity theories, counterfactual theories and manipulability theories are powerful theories about what causation is. Given how well supported each element of the physicalist triad is, if I have shown that they are also mutually reinforcing, perhaps this is just another reason to favour them. So why do I think we should try to break out of the physicalist triad?

The weakest point of the triad, or so it seems to me, is the account of agency it provides. The physicalist triad entails a physicalist/event-causalist description of agency and, as I explain in this chapter, this description of agency faces a number of problems. First, there is the problem presented by apparent counterexamples that involve deviant causal chains from mental cause to bodily movement. Second, there is the difficulty posed by the fact that sometimes agency is manifested through refrainment, i.e. by not doing anything. Third, there is the problem of giving an account of actions that are less than fully intentional. These problems will be familiar to anyone keeping track of debates within philosophy of action. However, I will argue that these problems are not three distinct issues. Instead they are all symptoms of a more fundamental issue with a physicalist/event-causalist description of agency, which is the assumption that the distinction between 'agential' and 'non-agential' can be understood in terms of a distinction between different kinds of event-causal sequence.

4.1 The disappearing agent

Very generally, agency refers to the power to act. Part of the task of philosophy of action is to explain what it is to act. The physicalist/event-causalist answer to this question construes what it is to act in terms of intentionality: what it is to act is to do something intentionally, which entails that all actions are intentional under some description. Davidson argued for this position, claiming that 'a man is the agent of an act if what he does can be described under an aspect that makes it intentional' (1971/2001a: 46).

What it is for an action to be intentional is then explained in terms of causation by a mental state of the agent, or a mental event involving the agent (this is the causal theory of action). The difference between a bodily movement that is intentional and one that it is not 'lies in the causal aetiology of what happens when a body moves' (Smith 2012: 387). And, according to physicalism, these mental items are realised by physical items—most plausibly, neural events, or perhaps physical events that are themselves complex and include neural events

as parts. The picture of human agency that emerges is a reductive one. What it is for a person to act is nothing more than the triggering of bodily movements by sub-personal events.

This picture of human agency is endorsed, at least partially, by Alfred Mele (1992a; 2003), who, in his own words, defends 'a causal perspective on intentional action' which consists of a pair of theses: '(1) all intentional actions are *caused* (but not necessarily deterministically so); (2) in the case of any intentional action, a causal explanation framed part in terms of *mental* items (events or states), including motivation-encompassing attitudes, is in principle available' (Mele 2003: 5). Mele says the second thesis can be developed by adding that 'the relevant mental items are realised in physical states and events that are important causes of intentional actions, and—owing to the particular relations of the mental items to the realising physical items, to appropriate counterfactual connections between the mental items and the actions, and to the truth of relevant psychological and psychophysical generalisations—the mental items properly enter into causal explanations of the actions' (2003: 5). Mele has also defended a 'causal approach to analysing and explaining actions', which he describes as the view that 'our actions are, essentially, events (and sometimes states, perhaps) that are suitably caused by appropriate mental items, or neural realisations of those items' (2000: 279). So, although Mele is focused primarily on defending the viability of a causal account of what intentional action is, he is at least open to the possibility that such an account could be given a physicalist development.

Berent Enç also defends the causal theory of action, which he describes as 'the proposition that an act consists of a behavioural output that is caused by the reasons the agent has for producing that behaviour—reasons that consist of the beliefs and desires of the agent' (2003: 2). As stated in Chapter 2, Enç believes that deliberation is a 'computational process' that results in an intention that in turn causes an item of behaviour. 'On this thesis,' states Enç, 'actions are defined as changes in the world that are caused by mental events.' Enç also states that he '[helps himself] to the assumption that mental attributes like beliefs, desires, hopes, value judgements and so forth are manifestations of the physical world, and that what is generally referred to as naturalism is the correct view about such mental events and states' (2003: 2). The physicalist/event-causal description of human agency is also defended, in whole or in part, by Brand (1984), Bishop (1989), Bratman (1987), Dretske (1988), and recently by Shepherd (2021).

Several arguments presented in philosophy of action appear to show that seeking to understand agency in terms of a distinction between different types of event causation cannot be done without misconstruing the agency concept. For example, Jennifer Hornsby (2004) argues that the physicalist/event-causalist description of agency 'leaves agents out', which is problematic because 'human beings are ineliminable from any account of their agency' (2004: 2).

This objection has come to be known as 'the disappearing agent problem'. According to this objection, an essential part of our concept of agency is that,

in acting, the agent herself brings about changes. However, as David Velleman (1992: 461) puts it, causal theories of intentional action entail that the agent is 'merely the arena' within which mental states or events cause bodily movements. The agent herself does not bring about what she intends. In this way, the agent 'disappears'. This cannot be right because a world where agents do not bring about the results of their actions is a world where there are no actions. This objection is, I believe, devastating to the physicalist/event-causalist description of human agency. However, it is often misunderstood.

One way it is misunderstood is to see it as begging the question against the event-causal theory of action. All versions of the event-causal theory of action hold that acting intentionally consists in the right kind of event being caused to happen, in the right way, by the right kind of mental antecedents. The core proposal of the event-causal theory is that acting intentionally is nothing over and above some special kind of event causation. The disappearing agent problem can seem like a straightforward denial of the event-causal theory's core proposal. The critic of the event-causal theory complains that the agent is missing from an account of her agency, while the event-causal theory's core thesis is that mental states causing bodily movements *is* the agent carrying out her agency. In her summary of the disappearing agent objection, Sarah Paul characterises the disappearing agent objection as committing a category mistake:

> The complaint is sometimes put in terms of the subject being a 'mere arena' in which psychological states are contained, such that she is not involved in the interactions between mind and body. But the Causal Theorist is in no way committed to this way of thinking about the relationship between the subject and her own mind. Indeed, this seems to be a prime example of a category mistake: 'I see that there are mental states, and a body that moves around in virtue of this mental activity, but where is the person that does the moving?' (2020: 56)

If the disappearing agent problem is understood this way, then event-causal theorists can respond by insisting that the agent does not disappear on their account because the agent's bringing about what she intends is identical with mental states of the agent causing bodily movements.

Another way the disappearing agent problem is misunderstood is to see it as revealing that the standard version of the causal theory of action—i.e. the version which says that an intentional action is a bodily movement which is caused by an intention to act, which is in its turn caused by desire for something and a belief about how to satisfy that desire—is insufficient to capture intentional agency. This is how Velleman (1992) understands the problem. Velleman argues that in the standard version of the causal theory of action there is nothing—no mental state, or causal sequence—that amounts to the agent taking an active part in her action. However, for Velleman, this does not show that no version of the causal theory of action can succeed. Velleman thinks the

disappearing agent problem shows that the causal theory of action needs to be modified but not rejected.

Velleman argues that the standard version of the causal theory of action actually succeeds as an account of what it is to act 'half-heartedly, or unwittingly, or in some equally defective way' (1992: 462). That is, the standard version of causal theory of action does capture *a kind* of action, but it 'describes an action from which the distinctively human feature is missing … not a human action par excellence' (1992: 162). Velleman's opinion is that sub-par action, which he describes as 'half-hearted', 'unwitting', 'defective', consists of mental states like desire, belief and intention taking our bodies from inactivity to activity. In cases of sub-par action, the flux of events—which includes mental events—operates through us but does not involve us—we play no active part. In human action par excellence, by contrast, we are involved and do play an active part.

> In a full-blooded action, an intention is formed by the agent himself, not by his reasons for acting. Reasons affect his intention by influencing him to form it, but they thus affect his intention by affecting him first. And the agent then moves his limbs in execution of his intention; his intention doesn't move his limbs by itself. The agent thus has at least two roles to play: he forms an intention under the influence of reasons for acting, and he produces behaviour pursuant to that intention. (1992: 462)

According to Velleman, the active part we play can be reduced to the causal role of some mental state of ours. Specifically, 'a motive that drives the agent's critical reflection on, and endorsement or rejection of, the potential determinants of his behaviour, always doing so from a position of independence from the objects of review' plays the functional role of the agent in action par excellence (1992: 477). As long as this higher-order motive is included in the event-causal story leading up to a bodily movement, the causal sequence described amounts to action par excellence. If the disappearing agent objection is understood as merely showing that the causal theory of action needs to be modified, then it does not disprove the physicalist/event-causal description of agency, nor does it give us a reason to break out of the physicalist triad.

The third way the disappearing agent problem is misunderstood is to see it as a problem for event-causal accounts of *a special kind* of action, as opposed to action in general. A key example of this kind of misunderstanding can be seen in Derk Pereboom's (2014) argument for understanding *free will* in terms of agent causation.

Pereboom argues that event-causal libertarian theories of free will are inadequate. Libertarians about free will believe that an action cannot be free if it is deterministically caused to happen by a prior event (incompatibilism). Pereboom argues that simply injecting indeterminism into the causal chain leading up to an action cannot secure freedom. This is because 'if only events are causes and the context is indeterministic, the agent disappears when it

needs to be settled whether the [action] will occur' (2014: 55). The point here is that, in a determinist event-causal sequence, prior events 'settle' whether an action occurs, therefore, for the incompatibilist, the action cannot be free, but in an indeterministic system *nothing* settles whether the action occurs, and for *that* reason the action cannot be free. Free action requires that *the agent* settles whether the action occurs or not. In an event-causal system, even one which involves indeterminacy, the agent is not settling anything—they have disappeared—and so free actions do not exist. The solution is to hold that an action is free just in case it is caused to happen by the agent. Now the agent, rather than any prior event, is the causal determiner of the action. The thought is that agent causation best captures the sense in which free agents need to, themselves, be the settlers of their actions.

Pereboom's argument fails as an argument against event-causal libertarian theories of free will. As Randolph Clarke (2017) argues, the event-causal libertarian can grant Pereboom's condition that an action is only free if the agent (and no prior event) settles whether the action occurs or not but insist that this condition is met on her account of free action. The event-causal libertarian can argue that whether the action occurs or not is settled by the agent when the action occurs. Prior to the agent's action it is an open question whether the action will occur or not. The matter is not settled prior to the agent's action because the events that cause the action do not deterministically cause the action. However, when the action occurs, the question of whether the action will occur or not is closed, and thereby settled. The occurrence of the action itself settles whether the action occurs. Therefore it is not the case that *nothing* settles whether the action occurs. The event-causal libertarian can argue that the occurrence of the agent's action *is* the agent's settling of whether the action occurs or not.

One might defend Pereboom's claim that, if only events are causes and the context is indeterministic, then the agent does not settle anything by insisting that the event-causal libertarian fails to specify conditions that are sufficient for the agent to do anything at all. If the agent does not act, then no event could constitute the agent's settling of something. However, this would change the target of Pereboom's argument. Pereboom explicitly accepts that it is still possible for agents to act even if only events are causes; he only argues that none of these actions can be free. To argue that the event-causal libertarian fails to specify conditions that are sufficient for the agent to do anything at all is a different argument. Pereboom's argument fails, I think, because the disappearing agent objection is really an issue about the possibility of action itself, it is not specifically to do with freedom.

I have presented three ways in which the disappearing agent should *not* be understood, so how *should* we understand this problem? The disappearing agent problem is not best expressed as a direct challenge to causal theories of action. So expressed, it can seem like it is begging the question. Furthermore, the disappearing agent problem is not about a special kind of action, e.g. action

par excellence or free action. The essence of the disappearing agent problem is that our general concept of agency is *fundamentally at odds with* a view of the world that assumes that causal reality is nothing but a chain of causally related events, a worldview where 'all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another' (Lewis 1986: ix). Perhaps the best expression of the problem comes from Abraham Melden:

> It is futile to attempt to explain conduct through the causal efficacy of desire—all that can explain is further happenings, not actions performed by agents. The agent confronting the causal nexus in which such happenings occur is a helpless victim of all that occurs in and to him. There is no place in this picture of the proceedings either for rational appetite or desires, or even for the conduct that was to have been explained by reference to them. (1961: 128–129)

Melden describes the aim of theories like the causal theory of action as 'futile'. He thinks that no event-causal theory of action could succeed; such a theory will always fail to adequately capture our thinking about agency. Melden claims that within 'the causal nexus' the agent becomes 'a helpless victim of all that occurs in and to him'.

This claim needs a bit of explaining. The issue is that when causal reality is viewed as nothing but chains of causally related events, everything in the causal world is something that *occurs* or something that *happens*. Occurrences and happenings are not things that anyone 'does'. So, when causal reality is viewed as nothing but chains of causally related events, the agent does not seem like an agent anymore, because the agent does not seem to do anything; the agent seems passive, like a victim. This metaphor of the agent becoming a 'victim' is why, I think, the disappearing agent problem can seem like it is about free action, or action par excellence, but it is important not to get carried away by the metaphor. The essential point is that there is something about our concept of agency and something about the idea of the causal world as consisting of nothing but chains of causally related events that do not marry: agency is about agents doing things—a causally related chain of events contains only what occurs or happens.

Thomas Nagel (1986) also expresses the disappearing agent problem well. For Nagel, part of the problem with the physicalist/event-causalist picture of agency is that there are important truths about agency that are lost when we view the causal world from a physicalist/event-causal perspective. On the physicalist/event-causal picture, the causal world is a 'spatiotemporal mosaic' of instantiations of categorical, objective properties (Lewis 1994: 474) or 'the fusion of all events throughout space-time' (Schaffer 2007: 83). According to Nagel, 'something peculiar' happens when we attempt to describe action from this 'objective or external standpoint'.

> Actions seem no longer assignable to individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part … There seems no room for agency in a world of neural impulses, chemical reactions, and bone and muscle movements. Even if we add sensations, perceptions, and feelings we don't get action, or doing—there is only what happens. (1986: 110–111)

Should the disappearing agent problem be taken seriously? Is our general concept of agency really *fundamentally at odds with* a view of the world that assumes that causal reality is nothing but a chain of causally related events? Is it really 'futile' to try to explain what it is to act in terms of causation of bodily movements by mental events? I think we should answer these questions affirmatively and, for me, this is the main motivation for breaking out of the physicalist triad. However, this is a bold claim and I will need to defend it.

The disappearing agent problem should be taken seriously because the physicalist/event-causal picture of agency fails in three important ways: it fails to solve the problem of deviant causal chain cases; it fails to account for refrainment; and it fails to account for the unity between intentional agency and non-intentional agency. The best explanation for these failures is because the physicalist/event-causal picture of agency leaves no room for the agent.

4.1.1 Deviant causal chains

The physicalist/event-causalist picture of agency construes what it is to act in terms of intentionality: what it is to act is to do something intentionally. The causal theory of action says that intentional actions are bodily movements caused, in the right way, by certain mental states of the agent or mental events involving the agent. In Chapter 2, I mentioned that the most significant source of disagreement about how the causal theory of action should be formulated concerns what constitutes *the right way* for a mental state or event to cause a bodily movement for there to be intentional action. Not just any causal chain from mental event to physical event is sufficient for there to be an intentional action. These mental states have to operate in the causal chain in the right way. A necessary condition for acting intentionally is that the agent is in control of what is going on with them. For there to be intentional action, the causal chain from mental item to bodily movement must be such that it constitutes the agent's control over their action. The causal chain cannot deviate from the kind of causal chain that occurs in a normal, uncontroversial case of intentional action. Davidson gives an example of a deviant kind of causal chain:

> 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, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally. (Davidson 1973/2001a: 79)

In this example, the climber has an end he wants to achieve, namely to rid himself of the weight and danger of holding the other man, and the climber reasons that loosening his hold is the best means to achieve this end. This belief–desire pair causes a bodily movement of a type that is rationalised by the belief–desire pair, just as causal theorists allege it would in an ordinary case of intentional action. But in this case the causal route from belief–desire pair to bodily movement involves an intermediary state of nervousness that 'robs the climber of control', as John Bishop (1989: 134) puts it. In this example, the climber did not let go intentionally. The challenge for the causalist that deviant causal chains present is to 'specify the sorts of causal paths that can count as the "right" way in which beliefs and desires must yield behaviour for genuine intentional action to occur' (Bishop 1989: 135).

There is great disagreement on what kind of causal chain from mental state to bodily movement is required for the agent to retain control over their action. Davidson himself doubted that a reductive analysis of intentional action could be developed from the idea that states of desiring and states of believing are causes of the actions they explain because of deviant causal chain cases. However, many have argued that a reductive causal analysis of intentional action is still possible, Davidson's nervous climber example notwithstanding. Davidson's example shows that the original causal theory failed to specify jointly sufficient necessary conditions for intentional action, but this does not mean that a more sophisticated version of the causal theory will also fail. Many more sophisticated versions of the causal analysis of intentional action have been offered since 1973.

One promising strategy is the 'sensitivity approach' (e.g. Bishop 1989; Mele 1992a; Mele 2003; Peacocke 1979). This approach suggests that a necessary condition for intentional action is that the bodily movement caused by the relevant mental state is 'responsive' or 'sensitive' to the content of the mental state. One way of spelling out this sensitivity requirement is in terms of counterfactuals: a bodily movement is sensitive to the mental state that caused it if and only if a slightly different bodily movement—one that conformed to the different mental state—would have occurred had the agent's mental state had a slightly different content.[15] Smith (2010) gives a clear example: suppose a pianist wants

[15] The counterfactual version of the sensitivity approach isn't the only version available. Peacocke (1979: 69) offers an alternative version. Peacocke argues that there is an intentional action if and only if the bodily movement is caused by an intention and that the intention differentially explains the occurrence of the bodily movement. A state or event differentially explains

to appear nervous to his audience and believes he can achieve this end by playing a C# instead of a C during his piece. The pianist's pressing C# is sensitive to this belief–desire pair if and only if the pianist would have pressed B had he thought that pressing B would achieve his goal. Cases of deviant causation are thought not to satisfy this sensitivity requirement.

However, this proposal faces a counterexample. Consider again my friend Amy, who has a device that can manipulate my brain and nervous system in the manner of the character Black from Harry Frankfurt's (1969) thought experiment. Amy can use this device to control my bodily movements as an engineer might control a remote-operated machine. When Amy uses her device, what happens with my body is not up to me; I am not in control of my bodily movements and therefore am not demonstrating agency. When Amy uses her device, she has taken control over what goes on with me. Now suppose that Amy uses her device to move my body to carry out my own intentions. For example, suppose I form the intention to make tea, and in response to this Amy uses her device to make me make tea. Suppose further that had I formed a different intention Amy would have used her device to make sure my body moved in conformity with my alternative intention.[16] In this strange example, the bodily movement that results from my intention to make tea is sensitive to the content of that intention. However, when Amy uses her device to manipulate my brain and nervous system, I am not performing an intentional action: I am not in control over what is going on with my body; Amy is. Bishop calls cases like this, where the causal path from intention to bodily movement passes through a benevolent second agent, 'heteromesial' causal chain cases.

A more recent suggested solution to the causal deviance problem, proposed by McDonnell (2015), also cannot deal with this counterexample. McDonnell suggests that there is an intentional action if and only if the mental cause of the bodily movement is 'proportional', in Stephen Yablo's (1992) sense, to the bodily movement. My intention to make tea is a proportional cause of my subsequent tea-making if and only if the following counterfactual conditionals are true:

1. Had my intention to make tea been absent, then I would not have made tea.
2. Had my intention to make tea been absent, then had I intended to make tea I would have made tea.

---

another when there is a law backing the explanation, according to which changes in the intensity or value of the explanandum are correlated (one-to-one) with changes in the intensity or value of the explanans. For the sake of brevity, I won't discuss Peacocke's version of the sensitivity approach here. See Sehon (1997) for a convincing argument that Peacocke's proposed criterion for intentional action is neither necessary nor sufficient.

[16] This counterexample is adapted from an example given by Peacocke (1979: 87).

These are both true even in the heteromesial case.

One obvious response to such cases is to stipulate that the causal chain cannot be heteromesial if intentional action is to occur. However, as Bishop points out, this cannot be right, as not every heteromesial causal chain is such that it blocks intentional action. Bishop (1989: 125) describes a case where machinery like Amy's is used to make sure that damaged neural pathways carry on functioning as normal (e.g. suppose some synapse isn't functioning properly; Amy's machinery might work by stimulating the second neurone when the first is in the right electrochemical state, just as the first neuron would if it were working properly). Even if Amy had to hold a switch down to keep the machinery working, so that the causal chain from intention to bodily movement must go via an action of Amy's, this would not necessarily mean that no intentional action is possible in this case. Suppose I'm the one with the damaged neural pathways, and Amy has to keep the machine switched on when I decide to make tea. In this case Amy is helping me carry out my intention to make tea by helping my nervous system remain in working order—she's an essential component of the causal chain that lets me carry out my intention, but it is less clear that I lack agential control in this case. It is not the involvement of a second agent *per se* that is adversative to intentional action but the manner of their involvement.

The problem posed by deviant causal chain cases may be solvable. It might be possible to give a counterexample-free specification of what constitutes the right path from mental cause to bodily movement for the bodily movement to count as an intentional action. On the other hand, the project of specifying what it is for a causal chain from intentions to bodily movement to be non-deviant may suffer a similar plight to that faced by the project of specifying necessary and sufficient conditions for knowledge, namely that every new proposal faces new counterexamples and the project seems nowhere near an end.[17] The pessimistic conclusion is that deviant causal chain cases should make us doubt that causation by a mental event constitutes what it is to act. In *some* cases, causation by a mental event seems to put the mental condition of a person in control of a bodily movement *at the expense of* the person themselves. If we cannot distinguish such cases from genuine cases of agency in event-causal terms, then the idea that personal control over one's body consists in causation of a bodily movement by a mental state is doubtful.

4.1.2 Refrainment

Sometimes human beings demonstrate their agency by not acting. For example, imagine I let a plant die by not watering it. This is an example of refraining from acting and thereby allowing something to happen. Other examples of

[17] See Zagzebski (1994) for an argument that Gettier-style counterexamples are inescapable for almost every analysis of knowledge.

refrainment include offending someone by not greeting them (Alvarez 2013: 104) or allowing a telephone to continue ringing by not answering it (Hornsby 2004: 5). Another interesting case comes from John Hyman (2015: 10–11). Hyman uses an example of a child being picked up by a parent to show that sometimes passivity is voluntary. With respect to being picked up, the child is passive, but being picked up is voluntary for the child. We know being picked up is voluntary for the child because the child could resist (e.g. by pushing away the parent or crying) but does not. We can further suppose that the child wants to be picked up and does not resist because she wants to be picked up. This qualifies the case as an instance of intentional passivity. In this case, the child is demonstrating an agential power, even though the child is, so to speak, not doing anything but, rather, letting something happen to her. The child is demonstrating agency by not resisting. In this case, there is an action, but it is the action of the parent not the child. In these examples, what occurs is at least partly up to the agent—the agents have that kind of control over what happens. This suggests that these examples are examples of agency. However, there are no actions in these examples, so their status as agential cannot be explained as the causation of an action by a mental state or event.

Bruce Vermazen (1985) describes a subclass of actions called 'negative acts'. One could challenge the claim that there are no actions in the examples above by arguing that in the examples the agents perform negative actions. However, I do not think the above examples are correctly described as negative acts. They are instead what Randolph Clarke (2014) calls 'omissions', which he argues are the absences of action. Maria Alvarez (2013) describes an example of refrainment that I think illustrates what a 'negative act' is, even though Alvarez would not herself describe the example as such. In Alvarez's example, an agent stands motionless in front of a laser-beam mechanism that controls a door and thereby prevents the door from closing by not moving. In this example, there is no positive performance by the agent. However, there does seem to be something that the agent does. Standing motionless seems to be an action, albeit one that is described in negative terms. In contrast, in the plant example, watering my plant is an action available to me that I simply do not do, thereby I allow other events (transpiration perhaps) to cause the plant to die. In the example where I offend someone by not greeting them, it is the absence of an act of greeting that matters. Similarly, in the telephone example, not answering the telephone is not an action negatively described but the absence of an action.

What makes refrainment a demonstration of agency? The causal theory of action is not equipped to answer this question. If being capable of agency is just to possess mental states that cause actions to happen, as the causal theory of action proposes, then it should be impossible for people to demonstrate agency when they do not perform an action. Because refraining is not acting, what makes refrainment a demonstration of agency cannot be expressed in terms of the mental causation of an action. However, examples of refrainment are not counterexamples to the causal theory of action. The causal theory of action is

only an account of *action*; it does not purport to explain what refrainment is. What these examples indicate is that the causal theory of action cannot tell the whole story about agency in terms of causation of an action by a mental event. To give a full explanation of what agency is, we need to explain why agency can be manifested not only by performing an action but also by refraining from acting, a fact that initially seems very puzzling given that agency is the power to act.

The important question is whether it is possible to explain refrainment in a way that abides by the physicalist and relationalist assumptions of the physicalist/event-causal picture of agency. Relationalism says that causation is always and everywhere a relation between distinct entities ('cause' and 'effect') that are normally supposed to be events. There may be some theories of what events are that allow something's not-happening to count as an event,[18] but on any theory that takes seriously the idea that events are happenings, this proposal that omissions are events is implausible: something's not-happening is not a thing that happens.[19] Clarke (2014) argues that omissions are non-entities; that is, they are not things that exist at all. Relationalism thus seems to rule out that omissions could be causes or effects. This appears to rule out any event-causal explanation of refrainment: the agency of refrainment cannot consist in omissions being caused to happen by mental states or events because omissions cannot happen.

Clarke (2014), however, offers an account of refrainment that appears to be compatible with relationalism about causation. Clarke argues that his account of refrainments is compatible with the view that omissions cannot be causes or effects as they are non-entities. Clarke argues that, 'in a case of intentional omission or refraining, relevant mental states (or events) must cause some of the agent's subsequent thought or conduct, even if they needn't cause the absence of some action' (2014: 75). As he puts it elsewhere, 'in cases of intentionally omitting or refraining, some intention with relevant content must play a causal role with respect to some of what subsequently does happen—with respect to one's subsequent thought and conduct' (2014: 78). For example, suppose we accepted that the child's desire to be picked up could not be the cause of her not resisting because not resisting is an absence and therefore cannot be an effect. Clarke's account of refrainment allows us to explain the intentionality of this omission as consisting in the child's desire causing some of the child's subsequent behaviour. Suppose wanting to be picked up caused the child to put her arms around the parents shoulders (and thereby make her being picked up easier)—that would be what makes the child's not-resisting intentional,

[18] For example, on certain theories of events as property exemplifications it might be possible for there to be negative events. Philosophers who have argued for the reality of negative events include: De Swart (1996), Higginbotham (2000) and Vermazen (1985).

[19] See Mele (2005) for further reasons to reject negative events.

according to Clarke's proposal. If Clarke's proposal succeeds, then the physicalist/event-causal account of agency only needs to be slightly amended to include refrainment in its account of agency. The amended account would be: what it is to demonstrate agency is to do *or not do* something intentionally and what it is for an action *or omission* to be intentional is explained in terms of causation by a mental state of the agent, or a mental event involving the agent.

Although Clarke's account of intentional omissions is similar to the causal theory of intentional action insofar as mental causation is an essential part of what makes an omission intentional, some of what Clarke says about intentional omissions is anti-relationalist in spirit. The intentionality of omissions, and hence the agency of omissions, does not consist in their being caused to happen by any event. Instead, what makes an omission intentional is that it sits within a larger sequence of thoughts and actions that demonstrates a teleological structure. To find what makes omissions intentional we must look at the wider context of the agent's behaviour. The intentionality of the omission is not revealed if we consider the omission in isolation. Instead we have to see the omission as part of a larger pattern of activity that is directed towards an end that is incompatible with performing the omitted act. Even if mental causation is essential for understanding the intentionality of refrainment, these cases lend support for the idea that the physicalist/event-causal picture of agency cannot be the whole story about agency.

4.1.3 Over-mentalisation of agency

The third important failure of the physicalist/event-causal picture of agency concerns its treatment of agency that is less than fully intentional. On the physicalist/event-causalist view, to act is to do something intentionally. Agency is thus explained in terms of intentionality. However, not all examples of agency are also examples of intentional action.

In what follows, I will give three examples of agency that lack the typical characteristics of intentional action. Most proponents of the physicalist/event-causal picture of agency assume that the typical characteristics of intentional actions are as follows:

(a) they are done for reasons, which is to say that there is a true description of the action which makes the action seem to the agent to be a sensible or rational or good thing to do;
(b) they are done in order to achieve a goal, which is to say that there is some further action that the agent is trying to complete and her intentional action is, she believes, a means by which she can complete that further action;
(c) they are subject to rationalising explanations, which is to say that they can be causally explained by facts about what the agent wants to do and

facts about what the agent believes about how to do it (note that, on the physicalist/event-causalist view, rationalising explanations are causal explanations).

Of course, I acknowledge that there are theories of intentionality where actions can be intentional even though they lack some, perhaps even all of these features. Therefore, there may be conceptions of intentionality under which the examples I give here *do* count as intentional. (Indeed, some of these examples may count as intentional on the conception of intentional action which I propose in Chapter 9.) However, as the purpose of this chapter is to challenge the physicalist/event-causal picture of agency, what matters dialectically is whether the examples I describe in this section conform to the physicalist/event-causalist characterisation of intentional action. If they do not, then the physicalist/event-causalist idea that to act is to do something intentionally is under pressure.

The first examples of agency that lack the typical characteristics of intentional action are actions that Brian O'Shaughnessy calls 'sub-intentional'. Sub-intentional actions include actions like 'tapping my feet to the music' and 'idly moving my tongue in my mouth' (1980: 61), actions we'd often describe as 'absent-minded'. Other examples may include shifting one's position, automatically scratching an itch or fiddling with one's hair.

Sub-intentional actions are not actions that seem, to the agent at the time of performing them, like sensible, or rational or good things to do, nor are they actions performed in pursuit of a goal. Sub-intentional actions also cannot be rationalised by facts about what the agent wants to do and what the agent believes about how to do it. Actions like tapping one's foot to music or shifting one's position, or fiddling with one's hair do not seem to be preceded by or accompanied by (and hence not causally explained by) an intentional state such as believing that performing the action is a good idea, or wanting to achieve something by means of the action. When I absent-mindedly tap my feet to the music, it is not true that I do this because there is something I want to do and believe that tapping my feet is a means by which I can do it. Furthermore, at the time of performing a sub-intentional action, the agent is often not aware that she is performing the action at all. The actions O'Shaughnessy delineates are actions about which we'd often say "Oh, I didn't realise I was doing that." For these reasons, sub-intentional actions, despite being under my control, seem to lack characteristics (a)–(c).

Sub-intentional actions also do not seem to be the causal consequence of an episode of thought. O'Shaughnessy thinks that sub-intentional actions are subject to psychological explanations. For example, he suggests that sub-intentional actions might be explained in terms of feelings of restlessness (1980: 61). When I shift my position, it is usually because I feel uncomfortable. I might fiddle with my hair because the sensation is comforting to me. I concede that sub-intentional actions can be explained in terms of feelings or sensations.

However, these psychological explanations do not seem to point to or mention a specific mental event that preceded the action and which could be considered the cause of the action. The explanations seem to cite concurrent experiences, as opposed to episodes in the agent's mental history, which caused her to fidget or fiddle. For this reason, sub-intentional actions seem to be exercises of agential power which do not have mental causes.

Another important class of human actions which lack the typical characteristics of intentional actions are *spontaneous expressions of emotion*. Examples include embracing a loved one, crying upon hearing bad news, laughing at a joke, wincing when you make a mistake, or shouting at your computer after it crashes at an inconvenient moment. Spontaneous expressions of emotion are distinct from reflexes like blushing when you are embarrassed or sweating when you are anxious. These reflexes seem entirely under the control of subpersonal systems. Spontaneous expressions of emotion on the other hand are behaviours that are up to us. Even when completely spontaneous, and so not preceded by any kind of conscious choice, they are still behaviours over which we are in control.

Like sub-intentional actions, spontaneous expressions of emotion are not actions we take for a reason: when we embrace a loved one or cry upon hearing bad news, we do not do these things because it is sensible or rational or good to do so. Such actions also do not seem to be accompanied by a desire to achieve a goal and a belief about how to achieve that goal. Of course, you *can* express an emotion in order to achieve something. For example, you might laugh at a joke not because you find it funny but in order to please the joke-teller. In this case, one could explain your laughing in terms of another activity you are engaging in, namely pleasing the joke-teller. However, examples like this are not truly spontaneous expressions of emotion. It is perhaps more accurate to describe them as emotional performances.

Furthermore, spontaneous expressions of emotion do not seem to be subject to rationalising explanations. Rosalind Hursthouse (1991) argues that spontaneous expressions of emotion cannot be explained by stating that the agent wanted to express an emotion (or vent it, or relieve it, or make it known) and believing that their actions constituted the expression of that emotion. Hursthouse correctly points out that many spontaneous expressions of emotion are simply not accompanied by a desire to express an emotion. Crying upon hearing bad news, for example, is often not something we want to do at all. Hursthouse also argues that it is wrong to suppose that the agent of a spontaneous expression of emotion possesses a belief about whether or not their behaviour constitutes an expression of the emotion they are expressing. The reason this would be wrong is because when an agent spontaneously expresses an emotion they *cannot be wrong* about whether what they are doing constitutes an expression of the emotion they are expressing. If I am crying to express my sadness, I cannot be wrong about whether my crying is an expression of sadness or not. Hence, it does not make sense to ascribe to me the belief that my crying

is an expression of sadness and to use this belief to explain why I am crying. This contrasts with actions which are subject to rationalising explanations. For example, when we explain why Carlin is adding rosemary to the sauce by stating that he wants to make the sauce taste better and believes adding rosemary will accomplish that, Carlin can be wrong about whether adding rosemary will make the sauce taste better. For this reason, it makes sense to ascribe to Carlin the belief that adding rosemary will make the sauce taste better and to use this belief to explain his action.

A third group of actions that do not display the typical characteristics of intentional actions is the actions of non-human animals. It is controversial whether non-human animals are capable of agency. We naturally speak of non-human animals doing things using the very same verbs we would use to describe some human actions: non-human animals hunt, seek shelter, raise young, climb, explore, cower, fight. However, in philosophy of action it is widely accepted that not everything an animal can be said to 'do' counts as an action of that animal. It is perfectly legitimate to speak of forgetting or falling over as things that one has done, even though forgetting and falling over are not, in any sense, actions. Reflex behaviours too can be things that we do—but they are not usually considered demonstrations of agency. There is a distinction between genuine actions, which are demonstrations of agency, and so-called 'mere behaviour': bodily movements that do not count as demonstrations of agency. It is controversial whether the bodily movements of animals count as agential or as mere behaviour.

One reason that philosophers have been reluctant to count the actions of non-human animals as demonstrations of agency, as opposed to mere behaviour, is because they have doubted that animals are capable of acting intentionally. Such arguments often rely on the assumption that animals lack the mental capacities that are prerequisites for intentional action.[20] For instance, it is doubtful that non-human animals are able to think of their actions as sensible or rational or good because non-human animals probably lack the ability to assess how well different courses of action could execute their intentions.

Many actions of non-human animals can be described as goal-directed. We often describe non-human animals as trying to do certain things. For example, the cat is trying to catch the mouse; the mouse is trying to hide. Furthermore, their behaviour demonstrates the kind of plasticity or flexibility we would expect if we assumed that they were acting in pursuit of a goal. If swiping towards the mouse's hiding place is unsuccessful, the cat might try waiting in ambush instead. Animals certainly seem to behave as if they were pursuing goals: they adjust their behaviour in response to changes in their environment, they change their behaviour to overcome obstacles, and they employ new tactics if their first attempts fail. However, Mele suggests that 'intentional action

[20] For example, Davidson (1982), Hacker (2007), McDowell (1996) and Stoecker (2009).

is not merely goal-directed action, but action directed in light of the agent's own goals, or desires; and desires, perhaps typically in conjunction with beliefs linking desired goals to prospective instrumental behaviour arguably constitute reasons for action' (1992b: 200). Even if animal behaviour is goal-directed, it is doubtful that non-human animals possess beliefs that link their goals to prospective instrumental behaviour.

We often successfully explain animal behaviour in terms of what the animal wants and believes. For example, the cat wants to catch the mouse; the mouse believes that under the sofa is a good hiding place. However, it is unclear that these explanations qualify as genuine rationalising explanations. Rationalising explanations, remember, explain why an agent acted as she did by telling us why, *in the agent's eyes*, what she did was a rational thing for her to do. It is not sufficient, then, that an attribution of a belief–desire pair makes the action intelligible *to us*. It is also necessary that the agent herself recognises that her desires and beliefs rationalise her actions. The agent needs to recognise that their action is desirable because it satisfies their own desire. It is at least questionable that non-human animals are able to do this.

The above examples contradict an important thesis which many supporters of the causal theory of action accept, which is that all actions are intentional under a description. The most obvious reply the causalist could make is to say that these examples are not really actions at all. The thought would be that sub-intentional action, spontaneous expression of emotion and the actions of non-human animals are not sufficiently distinct from passivity to qualify as actions at all.

Helen Steward (2009a) argues against this suggestion. She points out that it is completely natural to ascribe the production of the movements associated with sub-intentional actions to the person: 'when I fiddle with my jewellery … it is *me* who is fiddling with it, even if I am not aware that I am doing so' (2009a: 300). Steward thus has the opposite intuition to Velleman about these cases. Sub-intentional actions would count as *sub-par actions* for Velleman, and so they would be the kinds of actions for which Velleman is happy to say that the agent is not involved. Steward thinks the agent is very much involved in sub-intentional actions, and I agree. When I tap my feet to the music, it is me who is doing so. The movement is attributable to me as a person even if I am not performing this movement for the sake of any end or even with any awareness. Steward also emphasises that the agent of a sub-intentional action is *active* in bringing about the movements; it makes sense to speak of the person moving their body in these cases. To illustrate, consider how tapping your foot to music is very different to moving your foot because a doctor has triggered your patella reflex by tapping your knee. In the former case we would comfortably say that you are moving your body, even if you are not doing it on purpose; in the latter case we would say that your foot moved but not that you moved it. Furthermore sub-intentional actions are under the agent's control: 'The fiddling seems to be something which is under my *control*, and I seem to control it in

very much the same way that I control many of the processes which constitute my intentional actions (although in the sub-intentional case, the control is not exercised in the service of an end)' (2009a: 300). The agent of a sub-intentional action seems to possess exactly the kind of control over their movements that is lacking in deviant causal chain cases.

A similar argument can be made about spontaneous expressions of emotion and the actions of non-human animals. Spontaneous expressions of emotion are attributable to the person: no other agent, or sub-personal system, is acting through them. It is natural to speak of the person moving their body in cases of spontaneous expressions of emotion. And, even when completely spontaneous, expressions of emotion are behaviours over which we are in control. Another consideration that speaks in favour of counting spontaneous expressions of emotion as examples of agency is the fact that so much of our behaviour is emotionally driven. We might like to think that most of our actions are fully intentional, that most of our actions are done in pursuit of a goal, that we decide to do most of what we do, but I think that is wishful thinking. A great deal of what we do is done as an expression of emotion. In many situations, there simply isn't time to think about what to do before taking action. Choices are made, directions given, words spoken before any beliefs about the situations that called for those choices, directions or words have been formed. A lot of the time, we act spontaneously, using our feelings about a situation to guide us rather than our thoughts.

Similarly, at least some non-human animals seem capable of controlling their bodies in exactly the way that subjects in deviant causal chain cases cannot. As mentioned, many non-human animals display the kind of flexibility in their behaviour we would expect if they were acting in pursuit of a goal. They adjust their behaviour, they overcome obstacles, they try again if they don't succeed. This seems to imply that animals direct their own movements as opposed to passively undergoing changes in response to events occurring in their environment or inside their bodies. We also often speak of animals as moving their bodies and attribute their movement to them. As Steward argues elsewhere:

> It is most unnatural to suppose that the cockerel was caused to make its journey across the yard by anything like a mere reflex or a simple stimulus-response mechanism. For although we obviously have to recognise the huge importance of instinct in the lives of animals, instincts which prescribe for a given animal a range of basic activities from which it is certainly not free to forbear, I think we allow to the animal—and this is crucial, in my view, for the concept of agency—a certain freedom and control over the precise movements by means of which it satisfies those instinctual needs and desires. (2009b: 225)

Steward further defends her intuition that sub-intentional actions are genuine exercises of agential power by arguing against a line of thought that would

pull someone in the opposite direction. Her argument here could apply equally well to spontaneous expressions of emotion and animal action. Someone who wanted to discount these examples as actions might think that 'unless there is some reason to suppose that a movement is in some sense the product of something mental, there can be no reason to think it should be associated in any special way with the self, with the agent … Unless my mind is somehow involved, the thought goes, *I* could not be involved either' (2009a: 303). Steward argues that this thought stems from two prejudices.

The first is the Cartesian assumption that, if a person can control her own body, then the thing doing the controlling in that case must be the person's mind. Steward suggests that we think of some animals as being in possession of their bodies: some animals have bodies that they can to some extent control: 'We think and speak of animals—especially human ones—as *possessed of* their bodies, and to a certain extent, as controllers of them' (2009a: 303). However, this innocuous thought does not entail that, when an animal controls its movements, its mind controls its movements. Steward acknowledges that we typically attribute body-possession and mindedness together. It is an important truth that properties like having a mind, having thoughts and being conscious seem conceptually connected to properties like being the kind of creature that has a body it can control. However, Steward insists that the existence of this important connection does not entail that every time an animal controls its body this must be a case where the animal's *mind* controls its body. What is suggested by Steward's argument is that causal theories of action 'over-mentalise' agency by assuming that being able to control one's body entails the existence of a mind doing the controlling.

The second prejudice concerns the nature of causation. The assumptions about causation that Steward thinks prevent accurate appraisal of sub-intentional action, spontaneous expressions of emotion and animal actions are precisely the assumptions which constitute relationalism. Relationalism says that causal reality is nothing more than a chain of causally related events. This means that the causal truths about agency must be truths concerning causation of and by certain events; therefore, any distinction crucial to our conception of agency must be a distinction between different types of event causation. If one is committed to relationalism, then the idea of an animal controlling its movements *must* be reducible to a statement about an event occurring within the animal that produces the effect. This is why the distinction between intentional actions and other events becomes very important, because there is plausibility to the idea that mental causation is key to understanding *this* distinction. For Steward, this constitutes a *prejudice* because it forces us to think that, if non-intentional actions are actions at all, then they must have a mental cause instead of taking them at face value: genuine exercises of agential power that do not have mental causes.

Sub-intentional action, spontaneous expressions of emotion and the actions of non-human animals are not counterexamples to the causal theory of action.

They do not disprove the causal theory of action, as that theory is only intended to be a theory of *intentional* actions—it is not required to say anything about actions which are not intentional. However, any theory of intentional action should recognise the *continuity* between intentional action and other forms of agency. There is continuity between the kind of control demonstrated in non-intentional action and the kind of control demonstrated in intentional action. It would be wrong, I think, to say that these are examples of a completely different kind of control. Rather, intentional action is a *development* of the kind of agential control demonstrated in sub-intentional action, spontaneous expressions of emotion and animal action—it is the same fundamental phenomenon but extended or enhanced. The physicalist/event-causal picture of agency is poorly equipped to recognise this continuity as it ties the agency concept so closely to intentionality and mental causation.

4.2 Conclusion

I have presented three criticisms of the physicalist/event-causal picture of agency: it fails to solve the problem of deviant causal chain cases; it fails to account for refrainment; and it fails to account for the unity between intentional action and non-intentional agency. I now need to explain how these three failures connect to the disappearing agent problem.

The causal theory of intentional action aims to understand intentional action via a single divide: between event-causal sequences that involve intentional states and those which do not. However, the boundary between agential and non-agential does not map onto this divide. The two distinctions cut across each other. Sometimes a certain kind of mental causation is what stops an example counting as an instance of agency (deviant causal chain cases); our agency concept extends to cases where agents remain passive and so no bodily movement is caused to happen (refrainment); and our concept of agency extends to cases where there is no mental cause of a bodily movement (non-intentional action). What this suggests is that attempting to understand agency in terms of a distinction between event-causal sequences that involve intentional states and those that do not misconstrues the agency concept.

Common to all the diverse examples of agency described above is *the involvement of the agent*. In both Velleman's action par excellence, where an agent 'moves his limbs in execution of his intention', and sub-intentional action, where the agent is barely aware that they are moving their limbs and intentional states play no causal role, the agent is in control of their body—their bodily movements are up to them. When an agent refrains from doing something, the agent still retains some control over the situation in virtue of not exercising a power to act. And even if we cannot confidently say that animals act for reasons, or ascribe to them the propositional attitudes typically associated with intentional agency, they still seem to have control over the movement of their bodies.

The best explanation for why one cannot provide a comprehensive account of agency if one abides by the assumptions of the physicalist/event-causal picture of agency is because this picture leaves the agent out. If one assumes that causal reality is nothing more than a chain of causally related events, and therefore that the causal truths about agency are truths concerning causation of and by certain events, then any distinction crucial to our conception of agency must be a distinction between causal relations involving a mental relatum and causal relations that do not involve a mental relatum. However, this assumption leaves us unable to resolve the three issues described above. The distinction between agency and non-agency does not map onto a distinction between causation involving mental causes and causation not involving mental causes.

To adequately understand agency we need a metaphysical framework that allows us to see how the causality of action might be something that casts the agent herself as a causal player, rather than merely the setting for events to cause other events. The physicalist/event-causal picture of agency is unsatisfactory because our general concept of agency is fundamentally at odds with a view of the world that assumes that causal reality is nothing but a chain of causally related events.

References[21]

Alvarez, M 2013 Agency and two-way powers. *Proceedings of the Aristotelian Society*, 113(1pt1): 101–121. DOI: https://doi.org/10.1111/pash.2013.113.issue-1pt1

Bishop, J 1989 *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Brand, M 1984 Intending and acting. *Mind*, 96(381): 121–124.

Bratman, M 1987 *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Clarke, R 2014 *Omissions: Agency, metaphysics, and responsibility*. New York: Oxford University Press.

Clarke, R 2017 Free will, agent causation, and 'disappearing agents'. *Noûs*, 53(1): 76–96. DOI: https://doi.org/10.1111/nous.12206

[21] Author note: some references to Davidson are formatted (1963/2001a). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001a collection of his essays, with the page numbers relating to that volume. Similarly, some references to Davidson are formatted (1997/2001b) which indicates the initial date of publication (in this case 1997) but references the paper as it appears in the 2001b collection of Davidson's essays, with the page numbers relating to that volume.

Davidson, D 1971 Agency. In: Binkley, R, Bronaugh, R and Marras, A *Agent, action, and reason*. Toronto: University of Toronto Press. pp. 1–37. Reprinted in Davidson 2001a pp. 43–62.

Davidson, D 1973 Freedom to act. In: Honderich, T *Essays on freedom of action*. New York: Routledge and Kegan Paul. pp. 137–156. Reprinted in Davidson 2001a pp. 63–82.

Davidson, D 1982 Rational animals. *Dialectica*, 36(4): 317–328. Reprinted in Davidson 2001b pp. 95–106.

Davidson, D 2001a *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Davidson, D 2001b *Subjective, intersubjective, objective*. Oxford: Clarendon Press.

De Swart, H 1996 Quantification over time. In: van der Does, J and van Eijck, J *Quantifiers, logic, and language*. Cambridge: Cambridge University Press.

Dretske, F 1988 *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Enç, B 2003 *How we act: Causes, reasons, and intentions*. New York: Oxford University Press.

Frankfurt, H 1969 Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23): 829–839. DOI: https://doi.org/10.2307/2023833

Hacker, P 2007 *Human nature*. Oxford: Blackwell.

Higginbotham, J 2000 On events in linguistic semantics. In: Higginbotham, J, Pianesi F and Varzi, A *Speaking of events*. New York: Oxford University Press, pp. 49–81.

Hornsby, J 2004 Agency and alienation. In: Macarthur, D and De Caro, M *Naturalism in question*. Cambridge, MA: Harvard University Press. pp. 173–187.

Hursthouse, R 1991 Arational actions. *Journal of Philosophy*, 88(2): 57–68. DOI: https://doi.org/10.2307/2026906

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Lewis, D K 1986 *Philosophical papers*. New York: Oxford University Press.

Lewis, D K 1994 Humean supervenience debugged. *Mind*, 103(412): 473–490. DOI: https://doi.org/10.1093/mind/103.412.473

McDonnell, N 2015 The deviance in deviant causal chains. *Thought: A Journal of Philosophy*, 4(2): 162–170. DOI: https://doi.org/10.1002/tht3.169

McDowell, J 1996 *Mind and world*. Cambridge, MA: Harvard University Press.

Melden, A I 1961 *Free action: Studies in philosophical psychology*. London: Routledge & Kegan Paul.

Mele, A 1992a *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.

Mele, A 1992b Recent work on intentional action. *American Philosophical Quarterly*, 29(3): 199–217.

Mele, A 2000 Goal-directed action: Teleological explanations, causal theories, and deviance. *Noûs*, 34(14): 279–300. DOI: https://doi.org/10.1111/0029-4624.34.s14.15

Mele, A 2003 *Motivation and agency*. Oxford: Oxford University Press.
Mele, A 2005 Action. In: Jackson, F and Smith, M *The Oxford handbook of contemporary philosophy*. Oxford: Oxford University Press. pp. 78–88.
Nagel, T 1986 *The view from nowhere*. New York: Oxford University Press.
O'Shaughnessy, B 1980 *The will*. Cambridge: Cambridge University Press.
Paul, S 2020 *Philosophy of action: A contemporary introduction*. London: Routledge.
Peacocke, C 1979 Deviant causal chains. *Midwest Studies in Philosophy*, 4(1): 123–155. DOI: https://doi.org/10.1111/j.1475-4975.1979.tb00375.x
Pereboom, D 2014 *Free will, agency, and meaning in life*. New York: Oxford University Press.
Schaffer, J 2007 Causation and laws of nature: Reductionism. In: Sider, T, Hawthorn, J and Zimmerman, D W *Contemporary debates in metaphysics*. Malden: Blackwell, pp. 82–107.
Sehon, S 1997 Deviant causal chains and the irreducibility of teleological explanation. *Pacific Philosophical Quarterly*, 78(2): 195–213. DOI: https://doi.org/10.1111/1468-0114.00035
Shepherd, J 2021 *The shape of agency: Control, action, skill, knowledge*. Oxford: Oxford University Press.
Smith, M 2010 The standard story of action: An exchange 1. In: Buckareff, A A and Aguilar, J H *Actions: New perspectives on the causal theory of action*. Cambridge, MA: MIT Press. pp. 45–56.
Smith, M 2012 Four objections to the standard story of action (and four replies). *Philosophical Issues*, 22(1): 387–401. DOI: https://doi.org/10.1111/j.1533-6077.2012.00236.x
Steward, H 2009a Sub-intentional actions and the over-mentalization of agency. In: Sandis, C *New essays on the explanation of action*. Basingstoke: Palgrave Macmillan.
Steward, H 2009b Animal agency. *Inquiry*, 52(3): 217–231. DOI: https://doi.org/10.1080/00201740902917119
Stoecker, R 2009 Why animals can't act, *Inquiry*, 52(3): 255–271. DOI: https://doi.org/10.1080/00201740902917135
Velleman, D J 1992 What happens when someone acts? *Mind*, 101(403): 461–481. DOI: https://doi.org/10.7591/9781501721564-008
Vermazen, B 1985 Negative acts. In: Vermazen B and Hintikka, M B *Essays on Davidson: Actions and events*. Oxford: Clarendon Press. pp. 93–104.
Yablo, S 1992 Mental causation. *Philosophical Review*, 101(2): 245–280. DOI: https://doi.org/10.2307/2185535
Zagzebski, L 1994 The inescapability of Gettier problems. *Philosophical Quarterly*, 44(174): 65–73. DOI: https://doi.org/10.2307/2220147

CHAPTER 5

Agent Causation

In the previous chapter, I argued that we should try to break out of the physicalist triad since it provides an inadequate account of agency. The main failing of the physicalist/event-causal account of agency entailed by the physicalist triad is that it cannot provide a comprehensive account of agency—one that solves the problem of deviant causal chains, explains why refrainment counts as intentional action and accounts for the unity between intentional action and non-intentional action. The physicalist/event-causal account of agency is unable to deliver a comprehensive account because it leaves out the agent. This is the disappearing agent objection, and, although the objection is often misunderstood (see Section 4.1), I believe it is the most powerful objection against a physicalist/event-causal account of agency. The point of this objection is that our general concept of agency is fundamentally at odds with a view of the world that assumes that causal reality is nothing but a chain of causally related events. Thus, what is needed to adequately understand agency is a richer theory of causation, one that allows us to see how the causality of action might be something that casts the agent herself as a causal player, rather than merely the setting for events to cause other events.

Philosophers working within the field of philosophy of action and on the problem of free will have offered theories of what agency is which attempt to avoid the disappearing agent objection. Many of these accounts appeal to the notion of *agent causation*. According to this general type of view, agency is a kind of causation where the agent, who is taken to be a substance not an event, exercises causal power and this exercise of causal power cannot be reduced to causation by an event involving the agent. So, for example, what makes my action of typing this sentence a demonstration of agency is that *I* am causing letters to appear on my computer screen, where this *causing* of mine cannot be understood as the causation of one event by another (e.g. the causation of finger movements by a decision to type)—it is its own special type of causation.

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 97–114. York: White Rose University Press. DOI: https://doi.org/10.22599/White.f.

Appealing to agent causation to explain what agency is represents a departure from the standard relationalist view of causation, which takes all causation everywhere to be a relation between events or states. However, as we shall see, many agent-causationist accounts of agency accept some aspects of relationalism about causation and face significant issues as a result. In what follows, I will critically examine some agent-causal accounts of agency and argue that the chief failing of these theories is that they do not go far enough when it comes to rejecting relationalism about causation.

5.1 Traditional agent-causationism

I will first examine what I call 'traditional agent-causationism'. This title covers a family of theories which maintain that irreducible agent causation is required to adequately explain aspects of specifically human action. Traditional agent-causationists maintain that human agency is causation by an agent, who is taken to be a substance. According to this view, agent causation is a form of causation that cannot be identified with, or realised by, a causal relation between events or states. As such, causation by an agent cannot be analysed in terms of causation by any event involving the agent—causation by an agent is, in this sense, ontologically fundamental. An important tenet of traditional agent-causationism is that it is specifically human actions that must be understood as examples of irreducible agent causation. Traditional agent-causationists accept that *most* causation in the world, including interactions between non-human animals and inanimate objects, is nothing over and above causation of one event by another. It is only in the case of things done freely by human agents that there is something extra—causation by a *substance*.

Traditional agent-causationism is usually motivated by considerations to do with free will. Recall Pereboom's (2014) argument for understanding free will in terms of agent causation. For Pereboom, it is specifically *free* action that must involve irreducible agent causation because free agents need to, themselves, be the *determiners* or *settlers* of their actions. Roderick Chisholm also argues that agent causation is essential for an adequate treatment of *free will* (1976: 58–59). Richard Taylor similarly rests his case for an agent-causation-based account of agency on the idea that agents must be the 'initiators' or 'originators' of their actions (Taylor 1966: 112) and argues that this sense of 'initiation' or 'origination' is lacking in cases where inanimate objects cause things to happen. Taylor thus commits himself to the view that inanimate objects are never agents: 'a man is sometimes an agent who originates a change, and is not, like a match, merely a passive object that undergoes change in response to other changes' (1966: 122). Taylor denies that a match can be an agent because a match cannot 'wreak changes in itself': what a match does is always a response to the circumstances it is in and what's acting upon it. A person, in contrast, 'can bring about such a change as a motion of his arm quite by himself' (1966: 122). For this

reason, human agency must be understood in terms of an irreducible form of substance causation, but there is no similar demand to understand causation by inanimate objects in terms of irreducible substance causation.

Timothy O'Connor (2000; 2009) also argues that substance causation is a form of causation uniquely exercised by persons. O'Connor argues that 'an adequate account of freedom requires, in my judgement, a notion of a distinctive variety of causal power, one which tradition dubs "agent-causal power"' (2009: 230). In essence, O'Connor's view is that in order to make free choices about how we act, our actions need to be 'up to us'. Our actions are not up to us if prior events deterministically cause them, or so the thought goes. However, their being up to us also cannot consist in our actions being the non-deterministic causal consequence of certain events because, as O'Connor (2009: 231) puts it, 'looser connectivity in the flow of events' cannot constitute any kind of personal control over what happens. As O'Connor writes:

> [I]f I am faced with a choice between selfish and generous courses of action, each of which has some significant chance of being chosen, it would seem to be a matter of luck, good or bad, whichever way I choose, since I have no means directly to settle which of the indeterministic propensities gets manifested. (2009: 231)

The solution O'Connor endorses is to endow agents with a special causal power to bring about events, a power they must exercise if they are to act freely. In essence, the motivation for traditional agent-causationism is that agents themselves—and no events involving the agent—must cause their actions, otherwise free action is metaphysically impossible.

There are three key points to note about the metaphysics traditional agent-causationists think is required for agents to act freely. First, when an agent causes an event, the agent causes that event directly, which is to say that no event involving the agent or circumstances about the agent cause the event; in fact the event that the agent causes has no cause other than the agent herself. Second, agents cause their own actions. Third, there is no demand to understand causation by inanimate objects in terms of irreducible substance causation; in this way, human agency is metaphysically exceptional.

According to traditional agent-causationism, the event of my raising my arm, which is my causing my arm to rise, is an action because it is an event that I, *qua* substance, caused to happen. However, there is a well-known problem with this view. If my action is an event of which I am the cause, then we can ask of the causing of my action whether *this* is an action of mine or not. If it is, then, on the agent-causationist theory, it is also an event of which I am the cause, but now we seem to have opened an infinite regress: is the causing of my causing of my action another action? However, if we deny that the causing of my action is an action, then it seems we have two sorts of 'causings', some of which are actions and some of which are not. For example, my causing my arm to rise is

an action, but the causing of my causing my arm to rise is not—what makes this difference? It is unclear what the agent-causationist can or should say.[22]

Another important objection to traditional agent-causationism is that the account of agent causation I have just summarised is metaphysically unintelligible. Agent causation is supposed to be a special kind of causation distinct from and not reducible to a causal relation between events. The question is: what exactly is this special type of causation? What is it for an agent to 'directly cause an event' if this cannot be reduced to causation by an event involving the agent? Sceptics of agent causation argue that we have no independent understanding of what agent causation is. For example, von Wright (1971: 192) argues that the only way to make sense of agent causation is to see it as a synonym for human agency. This is especially the case if irreducible substance causation only exists in cases where a human being is acting freely. Limiting irreducible substance causation to exercises of free will makes substance causation seem like something discontinuous with the non-human world. Substance causation is made to seem like something additional to the world's normal causal functioning, which appears only when human beings act freely. Erasmus Mayr argues that:

> [R]estricting agent-causal activity to human agents (among the objects in the world) tends to make agent causation appear either as some unnatural extra force with which human beings are endowed and which can only be compared to divine causation—a comparison which is unlikely to improve our understanding of the notion—or as simply another name for the phenomenon we want to understand: human agency. (2011: 143)

The solution, which has been proposed by Mayr (2011) and Helen Steward (2012), is to insist that irreducible substance causation is ubiquitous. Non-human animals and inanimate objects cause things to happen in the same sense in which human beings cause things to happen when they act. In other words, substances causing things to happen is a general feature of the world, and human agency is just a special example of it. In no case can causation by a substance be reduced to causation by some event involving that substance. This may be because, in fact, all causation is fundamentally substance causation (Lowe 2008) or because causation is a diverse phenomenon and entities of many different categories—substances, events, facts, properties—can cause things, and each of these types of causation is fundamental (Hyman 2015; Mayr 2011; Steward 2012).

Even though the metaphysics proposed by traditional agent-causationists is ultimately unsuccessful, there are some aspects of traditional agent-causationism that I think are correct. What's right about traditional agent-causationism

[22] See Alvarez and Hyman (1998), Davidson (1971) and Hornsby (1980) for discussions of this problem.

is that the actions of humans (and I believe many non-human animals) are importantly metaphysically different from causation by inanimate objects. I think it is right to suppose, as O'Connor does, that human (and animal) agency must be a causal power of a special kind. The metaphysical exceptionalism of human (and animal) action is borne out in experimental philosophy. John Turri (2018) summarises findings from experiments that seem to suggest people think that 'human agency fits broadly within the causal order while still being exceptional in some respects' and more specifically that 'people believe human actions are caused by a variety of factors, including psychological, neurological, and social events' (2018: 402) and in that respect are part of the same causal order as everything else, but that humans and animals (but not computers or plants) were always capable of acting otherwise even if 'everything in the causal history of the physical world' rendered a certain outcome 100% probable (2018: 407). These findings do not tell us how to understand human agency, or how to spell out what is distinctive about it, but they demonstrate the pervasiveness of the intuition that human (and animal) agency is exceptional in some way.

I also think that traditional agent-causationists are right to seek to explain the metaphysical exceptionalism of human (and animal) agency by reference to the idea of things being up to the agent. I agree that it seems to be an essential part of our concept of agency that acting must involve a minimal kind of autonomy.

An essential characteristic of agency is that, when an agent acts, some of what goes on with the agent is up to the agent. One way to elucidate the idea of things being 'up to' the agent is to make use of Aristotle's distinction between self-movement and moved-movement. Humans and animals can move themselves; they do not need to be pushed or prodded or pulled by something else in order to move. Inanimate objects, on the other hand, can move, but they must be 'moved to move' by some other thing.

To help illustrate the distinction, consider the following examples. When a stone is thrown at a window with sufficient force, there is no sense in which it is *up to the stone* whether or not it breaks the window. If the conditions are right, i.e. the stone is heavy enough and the glass is thin enough, the stone will break the window (provided nothing comes along and interferes, e.g. no-one snatches the stone out of the air before it hits the window). The stone may well be the thing that is breaking the window—in this way the stone is a 'mover' (or, more precisely, a 'breaker')—but it was 'moved' to do so; that is, the stone was directed to break the window by some other thing (whatever threw it). Now consider the child who threw the stone. Ordinarily, when a child throws a stone the child moves his own body to move the stone. Even if the child was acting out of such intense emotion that we would not want to say the action was free, or intended, or chosen, no other thing is moving the child's arm for him. In this sense, the child is moving himself.

Robots that mimic human movements, such as Honda's ASIMO, are also moved-movers and not self-movers. This might seem counterintuitive because, unlike a stone, ASIMO can move around and perform various tasks without

another substance intervening. However, ASIMO's movements are strictly governed by his construction and programming. To illustrate: ASIMO has two cameras, a laser sensor, an infrared sensor and an ultrasound sensor. When information recorded by these sensors conflicts with information in ASIMO's pre-loaded map of navigable paths (e.g. by signalling that there is an obstacle in one of these paths), ASIMO cannot but move around the obstacle (American Honda Motor Co. Ltd. Public Relations Division 2007). ASIMO is moved to move around the obstacle by his component parts. It is not up to ASIMO what goes on with his legs. It is a necessary condition on our movements being up to us, and hence being genuine demonstrations of our agency, that we are not moved to move by our component parts.

Although I agree with traditional agent-causationists that there is an important metaphysical difference between human (and animal) action and causation by inanimate objects and that the metaphysical exceptionalism of human and animal agency has something to do with things being up to the agent in the former case but not the latter, I disagree that this metaphysical difference should be explained as the difference between two kinds of causation with different *ontologies*. I agree that human action is distinctive but I disagree that what makes human action distinctive is that it involves irreducible substance causation, whereas all other causation in the world is nothing over and above event causation. This is the wrong way to explain what makes human action exceptional. At best it makes human action seem like something unnatural. The capacity for self-movement is made to seem like a god-like capacity to directly interfere with event-causal chains. At worst it introduces a form of causation, causation of an event by an agent, which can only be understood as a synonym for human agency.

The reason writers like Pereboom, Chisholm, Taylor and O'Connor have gone wrong is, I think, because they have rejected relationalism about causation only in part. They accept the standard relationalist picture of causality with respect to animals and inanimate objects causing things to happen but reject it in the case of human agency. So, for example, when a stone breaks a window, the *real* cause is the event of the stone's being thrown towards the window, not the stone, but when a person breaks a window, the real cause is the person, not any event. However, this piecemeal departure from the relationalist picture is not justified.

One line of thought that might lead one to think that there must be substance causation in the case of human agency, but not in cases where inanimate objects cause things to happen, is as follows. Because inanimate objects are moved-movers, they are passive, which is to say they never cause change; they only suffer change. Therefore, the *real* cause in cases where an inanimate object makes something happen must be an event. However, this reasoning is fallacious. It is a fallacy to conflate moved-movement with passivity. Passivity is the manifestation of a passive power, or a liability, i.e. a power to undergo or suffer change. It contrasts with activity, which is the exercise of an active power,

i.e. a power to wreak change. Active powers are powers to change, and passive powers are powers to be changed. As John Hyman points out, the difference between agent and patient is not a difference between two different kinds of substance; it is rather a difference between two different roles substances can adopt (2015: 35). It is also possible for one and the same substance to be both agent and patient at the same time. For example, as Hyman notes, a victim of suicide is both agent and patient. Moved-movers when they cause change are both active and passive: active because they are causing a change but passive as well because their causing that change is dependent on another substance acting upon them.

A similar consideration that might lead one to think that there must be substance causation in the case of human agency, but not in cases where inanimate objects cause things to happen is discussed by Steward (2012). Steward considers the suggestion that, in cases where an inanimate object brings about an event, 'it is usually true that the object would not have caused the effect in question had it not been involved in some relevant event' (2012: 208). For example, the stone would not have broken the window had the child not thrown it. It would not have broken the window had it remained on the ground. From this, we may conclude that the event the stone is involved in is the real cause of the window-breaking. However, Steward argues that this reasoning is also fallacious. It depends on confusing causation and causal explanation. In order to adequately explain how the window came to be broken, we need to say something about how the stone came to break the window. It is rarely sufficient to answer the question 'why is the window broken?' by stating 'because of the stone'. However, as Steward points out, the fact that an adequate explanation requires reference to an event does not allow us to conclude that the stone 'does no causal work' (2012: 209). In Chapter 9, I will offer a positive account of how I think the crucial contrast between self-movement and moved-movement should be understood.

5.2 Actions-as-causings

The second agent-causation-based account of agency I shall consider does not contend that agents cause their own actions. According to this alternative agent-causation-based account, an agent's action *is* her causing of something; it is not what is caused. As Maria Alvarez and John Hyman put it, 'an action is a causing of an event by an agent' (1998: 224). I shall call this kind of theory the 'actions-as-causings' view. According to the actions-as-causings view, agency consists in an agent coming to stand in a causal relation to an event, or sometimes a state of affairs. However, what the agent causes is not her own action, instead the agent causes an event 'intrinsic' to the agent's action, an event that Alvarez and Hyman (1998: 233) call the 'result' of the action. The result of an action is not a causal consequence of the action; the relationship between

an action and its result is much tighter than that. For example, the result of an action of answering the phone is the event of the phone being answered. The action is what the agent does, and the result of the action is what must happen if the action is actually performed. Often, in the case of human action at least, the 'result' of an action is a bodily movement. For example, my action of raising my arm consists in my causing the rising of my arm. The rising of my arm is the result of my action and the event intrinsic to my action. *I* am the cause of my arm-rising, and my so being the cause of my arm-rising is what my action consists in.

The actions-as-causings view is most explicitly endorsed by Alvarez and Hyman (1998). However, Mayr also argues that human agency is an instance of substance causation (2011: 219), where substance causation should be understood in terms of a causal relation obtaining between a substance exercising an active power and the effect produced when the substance exercises active power: 'when such an "active power" is exercised, the cause of the resulting event is the substance which possess the power itself' (2011: 145–146). Similarly, E. J. Lowe describes agent causation as a species of causation 'in which the cause of some event or state of affairs is not (or not only) some other event or state of affairs, but is, rather, an agent of some kind' (2008: 121).

The crucial feature of the actions-as-causings view is that agency is described in terms of a causal relation, albeit one that obtains between an agent and an event or state of affairs. According to the actions-as-causings view, to properly understand agency we need to recognise that agents, *qua* substances, can be causes. The actions-as-causings view thus departs from standard relationalism insofar as it allows that substances can be relata of the cause–effect relation, not just events. However, substance causation is still described in relational terms. The action-as-causings view still accepts that causation is a relation between cause and effect; it just allows that substances—as opposed to only events—can be causes.[23]

My objection to the actions-as-causings view is that it entails two counterintuitive claims. First, the actions-as-causings view entails that one's actions are never identical to the bodily movements one's body makes when one acts. So, for example, my raising my arm cannot be identical with my arm's rising. Alec Hinshelwood calls this claim 'the separation thesis' (2013: 626). The second

[23] This kind of view is also endorsed by Harré and Madden (1975), who defend an account of causation as powerful particulars, which are substances, producing effects. For example, when a rock breaks a window, it comes to stand in a production relation to a window-breaking event. Thomas Reid also thought that causation was the production of change by the exertion of power and 'that which produces a change by the exertion of its power we call the *cause* of that change; and the change produced, the *effect* of that cause' (1788: 12–13).

counterintuitive claim the actions-as-causings view entails is that actions are not events.

For proponents of the actions-as-causings view, the separation thesis and the idea that actions are not events should not be seen as reasons to reject the actions-as-causings view. Instead they should be viewed as interesting, and inevitable, consequences of accepting that agency ought to be understood in terms of agent causation. However, this is incorrect. The separation thesis, and the idea that actions are not events, are not direct consequences of accepting that agency ought to be understood in terms of agent causation. Instead, these views are entailed specifically by the relational interpretation of agent causation endorsed by the actions-as-causings view.

5.2.1 Two counterintuitive claims

Alvarez and Hyman (1998) explicitly argue that actions are never identical to the movements one's body makes when one acts. Here is their argument:

> Davidson is one philosopher who claims that, in some cases, 'my raising my arm and my arm rising are one and the same event'. But my raising my arm is my causing my arm to rise. Hence, if my raising my arm is an event, it is the same event as my causing my arm to rise. And hence, if my raising my arm and my arm's rising are one and the same event, then my causing my arm to rise and my arm's rising are one and the same event. But it cannot be plausible that causing an event to occur is not merely an event itself, but the very same event as the event caused. (1998: 229)

Spelt out, the argument runs as follows:
Assume for *reductio*:

1. My raising my arm is one and the same event as my arm's rising.

Now assume the very plausible:

2. My raising my arm is my causing my arm to rise.

And:

3. If my raising my arm is one and the same event as my arm's rising, then my causing my arm to rise is my arm's rising.

Together these premises entail:

4. My causing my arm to rise is my arm's rising.

A conclusion that, when generalised, is revealed to be absurd:

5. My causing an event is the event caused.

In response to this argument, Alvarez and Hyman, and many writers sympathetic to agent-causation-based theories of action, have rejected the thesis that one's action is identical with the bodily movements one's body makes when one acts (the separation thesis).

To explain how the separation thesis is compatible with the plausible claim that many actions are bodily movements, Alvarez and Hyman (1998) make use of an ambiguity associated with the word 'movement' noted by Jennifer Hornsby (1980). Many verbs can be transitive (i.e. used with a grammatical object) or intransitive (i.e. used without a grammatical object). The verb 'move' is also ergative, which means that it can be transitive or intransitive and that the direct object of the verb when transitive becomes the subject of the verb when intransitive. For example, 'move' is transitive in the sentence "I moved my arm" but intransitive in "My arm moved", and the object of the transitive 'move' is the subject of the intransitive 'move'. This feature of the verb 'move' renders the nominalisation of 'move', 'movement', ambiguous. When we speak of, for example, my arm movement, there are two movements we might be talking about. There is one that corresponds to the transitive use of move, as in "I moved my arm", which can be otherwise picked out by the expression 'my moving of my arm', and the one that corresponds to the intransitive use of move, as in "My arm moved", which can be otherwise picked out by the expression 'the motion of my arm'. To help keep the two senses of 'movement' separate I will follow Hornsby's notation and use 'movement$_T$' for the first sense, and 'movement$_I$' for the second sense. Alvarez and Hyman (1998) hold that many actions are bodily movements$_T$, which they claim are *causings* of bodily movements$_I$, and hence cannot be identical with bodily movements$_I$.

Alvarez and Hyman (1998) also argue that actions, i.e. *causings* of bodily movements$_I$, are not events of any kind. To establish this conclusion, Alvarez and Hyman assume that there are only two possible sorts that event actions could be:

1. bodily movements$_I$; or
2. events that are causes of bodily movements$_I$.

Alvarez and Hyman take the first possibility to have been ruled out already by the argument outlined above. To show that bodily movements$_T$ are not events that cause bodily movements$_I$ Alvarez and Hyman argue as follows:

> [I]f bodily movements$_T$ are events which cause bodily movements$_I$, then either bodily movements$_T$ are events, perhaps neural events, which occur inside the agent's body, as for example Hornsby maintains in

> her book *Actions*, or they are events of another sort, which do not—presumably events which have no location at all, if there are such events. The first alternative implies that bodily movements$_T$, unlike their effects, are not normally perceptible without a special apparatus. The second implies that bodily movements$_I$ are caused both by neural events and by events of another sort, and therefore raises the difficult question of how these two sorts of events are related. It also implies that bodily movements$_T$ can never be perceived, whatever sort of apparatus we are equipped with. But we can and do see people and animals moving their limbs without making use of any sort of apparatus; and seeing a person or an animal moving its limbs is seeing a bodily movement$_T$. Hence neither alternative is tenable; and it follows that bodily movements$_T$ are not events which cause bodily movements$_I$. (1998: 229–230)

I agree that the first option Alvarez and Hyman consider here, that all actions are events that take place inside the agent, is not very plausible. Common sense suggests that many actions are public, and actions that involve moving one's body are paradigm examples of actions that other people can see without any special equipment. The second option Alvarez and Hyman consider is not as obviously implausible, partly because the option they suggest is itself difficult to understand. Explained with an example, the suggestion is that my action of raising my arm—which is assumed to be my causing of my arm's rising—is an event that causes my arm's rising but is not identical with any neural event or muscular event or indeed any of the events that occur in the vicinity of my arm's rising that are causally linked to my arm's rising. Instead it is an event that causes my arm's rising but is not located anywhere in particular. Put this way, the suggestion is very strange and Alvarez and Hyman are right to reject it.

If Alvarez and Hyman's argument succeeds, then bodily movements$_T$ are not events, so the causing of an event by an agent is some *other* sort of entity. The actions-as-causings theory of agency thus seems to involve ontological commitment to a novel kind of entity, which is the coming-to-obtain of a causal relation between an agent and an event. To give these novel entities a name, let's call them 'causings'.

Alvarez and Hyman (1998) are not the only philosophers who argue that actions are not events. This idea has quite a long history. Kent Bach (1980) argues that actions are not events because they are the obtaining of a causal relation between an agent and an event (see also von Wright 1962 and Chisholm 1964). There is some intuitive plausibility to the idea that actions are not events because actions can be said to be things people do and you cannot 'do' an event—an event is something that happens. However, this intuition is not robust enough to support a metaphysical conclusion because the word 'action' can be used in many different ways. Often the word is used to name activities people engage in—things people do—and in that sense does not seem to refer to a set of events. For example, "She took decisive action" probably refers to

the deeds the agent undertook—the things she did—and not the events that happened. However, there are many other uses of the word where it is more plausible to assume one is talking about things that happen, i.e. events. For example, "The action surprised her" could plausibly be interpreted as referring to something that happened. Similarly, "His action triggered a revolt" also seems to reference an event.

My own view is that the idea that actions are events is, to borrow an expression from Hornsby (2004), an *innocent* one. There is nothing majorly wrong with the idea that actions are events. Although the claim that actions are events is a key claim of event-causal theories of action, it is not the claim that does the most damage to our understanding of agency. Event-causal theories of action fail to adequately explain agency because they assume causal reality is nothing but a chain of causally related events and hence that what it is to act reduces to causal relations between events. Thus, the claim that does the most damage is not that actions are events but that what makes something an action is a question of what causal relations it is involved in. The best account of agency would be one that allows that *sometimes* when we talk about actions we are talking about events because that is what our language seems to imply.

I also think that the idea that one's actions are, at least sometimes, identical to the bodily movements one's body makes when one acts is similarly innocent. There is nothing majorly wrong with the idea that my raising my arm and my arm's rising are one occurrence. Indeed, I find the separation thesis counterintuitive for two reasons.

My first reason comes from an argument against the separation thesis made by Hinshelwood (2013). Hinshelwood argues that the separation thesis generates two epistemological issues, the first of which seems to me the most pressing. Hinshelwood begins his argument by pointing out that 'we can perceive what someone is doing simply by *seeing her doing it*' (2013: 628). In other words, actions are direct objects of perception—we can literally and directly see actions. For example, when someone raises their arm we do not see something else that serves as visual evidence of their action; we see the action itself. Hinshelwood then argues that the separation thesis calls this apparent epistemological datum into doubt. It is undeniable that the motions of people's bodies are directly visible. If, as the separation thesis claims, the movements one's body makes when one acts are not identical with one's actions, but are instead the results of one's actions, 'then we might be unsure whether we really can literally see the action itself'. If someone's arm rising is not their arm raising, then '[w]hat else could one see, the seeing of which would count as one's having seen the action?' Hinshelwood answers that 'there is nothing else available for one to perceive' (2013: 629–630). As I understand it, the problem that Hinshelwood identifies is that it is difficult to understand how the following statements can all be true: (a) we can directly see actions, such as someone's raising their arm; (b) we can directly see the bodily movements that are the results of actions, such as someone's arm going up; (c) according to the separation thesis these two things are not one

and the same. Hinshelwood thinks that the upshot is that we end up doubting that the action is really directly visible after all.

Hinshelwood acknowledges that there are other examples where we can directly see two objects that are visibly indistinguishable but nevertheless distinct. The most famous case is that of a bronze statue and the lump of bronze from which it is made. The statue and the lump are visibly indistinguishable; nevertheless, when we look in their direction we are looking at two objects, not one. The statue and the lump must be distinct objects because they each have different modal properties. The statue cannot survive being melted down, whereas the lump can. By Leibniz's law, if X and Y have different properties, then X and Y are not identical. In this case, we do not doubt the visibility of either the statue or the lump. It is not puzzling to say that there are two visibly indistinguishable, spatiotemporally coincident objects and both are directly visible because when we see one we are seeing the other.

Why, then, does Hinshelwood think it is puzzling to make a similar claim about actions and the movements one's body makes when one acts? Why can we not simply say that there are two visibly indistinguishable, spatiotemporally coincident eventualities (an action and a bodily movement$_I$) and both are directly visible because when we see one we are seeing the other? Hinshelwood's answer is because the two cases are not exactly analogous, and hence the action case can be puzzling even while the statue–lump case is not. In the statue–lump case we understand how the statue and lump can both be directly visible even though they are distinct by explaining that the lump *constitutes* the statue. It is understandable how we see one when we see the other because the one *constitutes* the other. If we wanted to explain how it is that someone's action and the movement their body makes when they act are both directly visible even though they are distinct, we would have to posit a relation similar to constitution to underpin their spatiotemporal coincidence and visual indistinguishability. Hinshelwood argues that it is doubtful that a relation of constitution holds between actions conceived of as causings and bodily movements$_I$ as the latter are supposed to be the causal results of the former.

Helen Steward (2013) offers a counterargument. She argues that, actually, it is not constitution that helps us understand how the statue and lump are both directly visible despite being distinct. What does the explanatory work here, according to Steward, are facts about how we individuate things. We understand that the statue and lump cannot be one and the same because of Leibniz's law. Because the statue and the lump have different modal properties, we understand that they cannot be identical. Actions and bodily movements$_I$ also have different properties, Steward suggests. Actions are things that are done; bodily movements$_I$ are not. Actions can be, for example, eager; bodily movements$_I$ cannot. This is sufficient to explain how actions and bodily movements$_I$ can be distinct even though they are visually indistinguishable.

However, I do not think this reply succeeds. This is because the puzzle that needs explaining is not how two things can be distinct despite being visually

indistinguishable. The puzzle is *how two things can both be directly visible* if they are distinct. What needs explaining is how when we see one we see the other. It is not enough to be reassured that the two objects are really distinct despite their visual indistinguishability. We need some explanation of what underpins their visual indistinguishability that we can use as reassurance that we really can directly see them both. It seems to me that it really is constitution, and not Leibniz's law, that explains that puzzle.

Thus, one reason to doubt the separation thesis is that it opens up a challenge to explain how it is possible that actions and bodily movements$_1$ can both be directly visible even though they are distinct. Of course, this puzzle may be solvable—just as it is in the statue–lump case. Even if constitution is not the right way to solve it, proponents of the actions-as-causings view may be able to give some other account of the relation between actions and bodily movements$_1$ that makes it clear how when see one we see the other.

The second reason I find the separation thesis counterintuitive is because of what it seems to imply about our relationship to our own bodies. Adrian Haddock (2005) suggests that, if the separation thesis is true, then persons are alienated from their bodily movements. According to Haddock, if the separation thesis is true, then 'our bodies are pictured as entities whose powers are wholly distinct from our powers of agency, as entities that we can (at best) only cause to move—and in this respect they are the same as any other worldly object' (2005: 161). I am not sure the separation thesis entails something quite as strong as that. The separation thesis does not, for instance, entail that moving my body is not a basic action. It does not entail that in order to move my body I must first do something else, as I have to do when I want to move other worldly objects: to move them, I need to move my body first. The separation thesis does not, therefore, collapse this distinction between moving our bodies and moving other worldly objects. Furthermore, on the view we are currently considering, actions are causings of bodily movements—they are not events that are the causes of bodily movements. This means that there is a difference between moving a glass of water and moving my arm in order to pick up a glass of water because in the first case we could say that the movement of the glass of water is caused by a prior event that I cause, i.e. the movement of my arm, whereas in the second case we cannot say that the movement of my arm is caused by a prior event that I cause, because the movement of my arm *is* what I cause. Therefore, I do not think it is correct to say that the separation thesis entails that bodies are treated 'the same as any other worldly object'.

However, I agree with the general discomfort Haddock expresses. The separation thesis says that what happens with my body when I act is not my action; it is instead the result of my action. This, to me, implies that my action constitutes my executive supervision, as it were, of what goes on with my body. What goes on with my body would not happen without me—I am the cause of my bodily movements$_1$, after all—but I am somewhat pulled back from what is happening with my body.

Indeed, Steward describes the control agents have over their own bodies using the metaphor of a supervisor (2012: 51, 52, 68, 162, 165). However, I do not think that we are present in our bodies as supervisors. To me, the separation thesis has parallels with Cartesian dualism. Instead of thinking of ourselves as one thing that is both physical and capable of apparently non-physical activities such as thinking, Descartes concluded from his meditations that we must be two separate substances joined together: a body and a mind. Descartes posited an additional entity—a mind—to be that which thinks, rather than accept that some physical things might be capable of non-physical activities. The separation thesis strikes me as similar in some ways. The separation thesis posits an additional entity—a causing—to be the exercise of our agential power, rather than accept that some bodily events are exercises of our agential power. Also, like Descartes's mind–body distinction, the separation thesis distinguishes our agency into a personal and bodily aspect. I am uncomfortable with this distinction between ourselves and our bodies. I agree with Haddock that the powers of our bodies are not wholly distinct from our powers of agency. We have the agential powers that we have only because of what our bodies are capable of. For example, I can lift things because of the power of my brain to stimulate my muscles and the power of my muscles to move my bones etc. My intuition is that the connection between ourselves and our bodies is much closer than that of supervisor and supervisee.

5.2.2 A response to Alvarez and Hyman

One way to prove the innocence of both the separation thesis and the idea that actions can be events is to show that, actually, both ideas are consistent with accepting that agency ought to be understood in terms of agent causation. I think that Alvarez and Hyman's argument for the separation thesis and for the conclusion that actions are not events is invalid. Alvarez and Hyman's argument is invalid because it wrongly assumes that the expression 'caused to rise' means 'caused an arm-rising event to happen'. Alvarez and Hyman rightly claim that it is implausible 'that causing an event to occur is not merely an event itself, but the very same event as the event caused' (Alvarez & Hyman 1998: 229). However, this only falsifies the claim that my causing my arm to rise is my arm's rising if 'causing my arm to rise' is taken to mean 'caused an arm-rising event to happen'. But why should we 'relationalise' the infinitival phrase 'causing my arm to rise'? Why should we assume that what claims like 'the agent caused her arm to rise' mean is that an agent is the cause of an arm-rising event? Rowland Stout rightly points out that '[t]he phrase "your arm to rise" is not really a noun phrase at all and certainly does not encode some implicit reference to an entity which is the event of your arm's rising' (2010: 104). In other words, the *language* we use to talk about what an agent causes when they act does not entail the metaphysical conclusion that when an agent raises her arm, a relation

of causation comes to obtain between the agent and an arm-rising event. The thesis that agency consists in an agent coming to stand in a causal relation to an event is a substantive metaphysical thesis—it is not simply what phrases like 'the agent caused her arm to rise' mean.

Ursula Coope (2007) outlines a response to Alvarez and Hyman's (1998) argument that is available to Aristotle, who also thought that my arm's going up, the arm-rising event, was identical with my action of raising my arm. Coope suggests that Aristotle would deny that his view commits him to the implausible idea that the causing of an event is one and the same as the event caused, because Aristotle would deny that an action is a causing of an event to happen. According to Coope's Aristotle, an action is the causing of a state to obtain:

> Aristotle's view, I shall argue, is that the power that is exercised in an action of moving X is a power to produce the end of X's movement: a power to produce a state, rather than a movement. In this sense, what I am causing when I move X is the state that X's movement is directed towards. For example, when I raise my arm, what I am causing is my arm's being up, rather than my arm's going up. More generally, the action of changing something towards being F is, for Aristotle, a particular kind of causing of the state being F. (2007: 113–114)

However, another more radical response to Alvarez and Hyman is available. Suppose we rejected the relational interpretation of 'causing my arm to rise'. Suppose we thought that an agent's causing her arm to rise does not entail that the agent stands in a causal relation to anything. So not only does an agent raising her arm not stand in a causal relation to an arm-rising event; she also does not stand in a causal relation to the state of her arm being up. Now it is possible to accept that my causing my arm to rise is my arm's rising, and that actions are events, because this no longer entails the absurd claim that my causing an event to happen is the event caused.

The temptation to assume that claims like 'the agent caused her arm to rise' mean an agent is the cause of an arm-rising event is a consequence of an incomplete rejection of relationalism. Relationalism says that causation is always and everywhere a relation between distinct entities ('cause' and 'effect'). Those who endorse the actions-as-causings view reject an event-causal theory of agency and so reject the idea that what it is for a person to act can be analysed in terms of some kind of relation between two events. However, they still seek to explain agency in terms of a causal relation. Agent causation is understood in relational terms: it is taken to be a relation of causation that obtains between an agent and an event. As a consequence of this partial rejection of relationalism, actions are construed as the-coming-to-obtain of a causal relation between an agent and an event *and not* as events themselves. The positive view I will advance in the following chapters involves the complete rejection of relationalism, which

allows me to retain much of what seems right about the actions-as-causings view, without also having to accept the separation thesis or that actions are not events.

References

Alvarez, M and Hyman, J 1998 Agents and their actions. *Philosophy*, 73(284): 219–245. DOI: https://doi.org/10.1017/s0031819198000199

American Honda Motor Co. Ltd. Public Relations Division 2007 ASIMO Technical Manual, September 2007. Available at https://asimo.honda.com/downloads/pdf/asimo-technical-information.pdf [Last accessed 3 September 2023].

Bach, K 1980 Actions are not events. *Mind*, 89(353): 114–120. DOI: https://doi.org/10.1093/mind/lxxxix.353.114

Chisholm, R 1964 Human freedom and the self. In Kane, R *Free will*. Malden, MA: Wiley Blackwell.

Chisholm, R 1976 *Person and object: A metaphysical study*. London: Routledge.

Coope, U 2007 Aristotle on action. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 81: 109–138. DOI: https://doi.org/10.1111/j.1467-8349.2007.00153.x

Davidson, D 1971 Agency. In: Binkley, R, Bronaugh, R and Marras, A *Agent, action, and reason*. Toronto: University of Toronto Press. pp. 1–37. Reprinted in Davidson 2001 pp. 43–62.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Haddock, A 2005 At one with our actions, but at two with our bodies: Hornsby's account of action. *Philosophical Explorations*, 8(2): 157–172. DOI: https://doi.org/10.1080/13869790500095939

Harré, R and Madden, E H 1975 *Causal powers*. Oxford: Blackwell.

Hinshelwood, A 2013 The metaphysics and epistemology of settling: Some Anscombean reservations. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 625–638. DOI: https://doi.org/10.1080/0020174x.2013.841044

Hornsby, J 1980 *Actions*. London: Routledge.

Hornsby, J 2004 Agency and actions. In: Steward, H and Hyman, J *Agency and action*. Cambridge: Cambridge University Press. pp. 1–23.

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Lowe, E J 2008 *Personal agency: The metaphysics of mind and action*. New York: Oxford University Press.

Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.

O'Connor, T 2000 *Persons and causes: The metaphysics of free will*. New York: Oxford University Press.

O'Connor, T 2009 Agent-causal power. In: Handfield, T *Dispositions and Causes*. Oxford: Oxford University Press.

Pereboom, D 2014 *Free will, agency, and meaning in life*. New York: Oxford University Press.

Reid, T 1788 *Essays on the active powers of man*. Edinburgh: John Bell, Parliament-Square, and London: G G J & J Robinson.

Steward, H 2012 *A metaphysics for freedom*. Oxford: Oxford University Press.

Steward, H 2013 Responses. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 681–706. DOI: https://doi.org/10.1080/0020174x.2013.841055

Stout, R 2010 What are you causing in acting? In: Aguilar, J H and Buckareff, A A *Causing human actions: New perspectives on the causal theory of action*. Cambridge, MA: MIT Press.

Taylor, R 1966 *Action and purpose*. Englewood Cliffs, NJ: Prentice-Hall.

Turri, J 2018 Exceptionalist naturalism: Human agency and the causal order. *Quarterly Journal of Experimental Psychology (Hove)*, 71(2): 96–410. DOI: https://doi.org/10.1080/17470218.2016.1251472

von Wright, G H 1962 On promises. *Theoria*, 28(3): 277–297.

von Wright, G H 1971 *Explanation and understanding*. Ithaca, NY: Cornell University Press.

PART 2

A Non-relational Understanding of Mental Causation

CHAPTER 6

A Non-relational Approach to Causation

In Chapter 4, I explained my reasons for wanting to break out of the physicalist triad. I argued that the triad provides an inadequate account of agency. I argued that physicalist/event-causal theories of agency are unable to deliver a comprehensive account of agency because they leave the agent out. This is the disappearing agent objection, which essentially claims that our general concept of agency is fundamentally at odds with a view of the world that assumes that causal reality is nothing but a chain of causally related events. Thus, what is needed to adequately understand human agency is a richer theory of causation, one that allows us to see how the causality of action might be something that casts the agent herself as a causal player, rather than merely the setting for events to cause other events.

In the previous chapter, I examined some existing alternatives to physicalist/event-causal accounts of agency that attempt to avoid the disappearing agent objection. Many of these accounts appeal to the notion of *agent causation*, a kind of causation that cannot be reductively analysed in terms of a causal relation between events. According to this general type of view, agency is a kind of causation where the agent herself exercises causal power and this exercise of causal power cannot be reduced to causation by an event involving the agent. I argued that the chief failing of existing agent-causation-based theories of agency is that they do not go far enough when it comes to rejecting the relational approach to causation. Existing agent-causation-based theories of agency do not cleanly break out of the physicalist triad and suffer problems as a result.

In the remainder of this book I will show how broadening our understanding of causation, and more specifically incorporating the concept of *process* into our understanding of causation, opens up new ways of understanding intentional action and the mental causation associated with it. In this chapter, I present my own non-relational approach to causation.

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 117–140. York: White Rose University Press. DOI: https://doi.org/10.22599/White.g.

6.1 Rejecting relationalism

A theory of causation is relational if and only if it is committed to the following thesis:

> **Relationalism:** causation is always and everywhere a relation between distinct entities ('cause' and 'effect'); the worldly phenomenon that is referred to by our concept 'causation' is not ontologically diverse in this respect.

One important feature of relationalism is that 'cause' is an unequivocal term. All causation everywhere is the same, so the only thing that can discriminate between different categories of causation is the nature of the relata involved.

A non-relationalist approach to causation denies that causation is always and everywhere a relation between distinct entities. One way to be a non-relationalist about causation is to deny that causation is ever a relation, and maintain that causation is something else instead. One theory of causation that I think does this is proposed by Steven Mumford and Rani Lill Anjum (2011). Mumford and Anjum proposed a powers-based theory of causation. On a powers-based theory of causation, facts about what powers things have, or what things can do, cannot be analysed as claims about what events regularly follow on from what others. On powers-based theories of causation, just like on realist theories of causation, causation is something in nature that constrains the ways in which events can unfold, and which therefore *grounds* regularity. In other words, worldly events unfold in a regular way *because* causation exists. On a powers-based theory, causation is the exercise of power, and worldly events unfold in a regular way because what can occur is limited by what powers entities possess: an entity with certain powers must behave in certain ways when the conditions for the manifestation of the power arise, provided there is nothing interfering with the entity and thereby blocking the manifestation.

Steven Mumford (2009) argues that no powers-based account of causation can be reductive, because *power* is a causal notion. For example, it is impossible to understand what it is to have the power to intoxicate without having some grasp of the phenomenon of intoxication, which is a causal process. Mumford (like Woodward (2003)) insists that an account of causation can be informative without being reductive. That is, an account of causation can give some insight into the nature of causation without telling us what non-causal structures exhaustively constitute causation. However, given that the powers-based theory takes causation to be the exercise of power, without saying more about what an exercise of power is, this account is in danger of seeming uninformative, perhaps even circular. What is missing from powers-based theories of causation is a suitable ontology that tells us what an exercise of a power is, what sorts of entities possess and exercise powers, and what sorts of relations those things stand in when they exercise their powers. In the rest of this section, I

briefly outline the ontology that Mumford and Anjum offer before offering in Section 6.2 what I believe to be a more satisfactory alternative.

Mumford and Anjum claim that 'the world is a world containing real powers' (2011: 4). In other words, Mumford and Anjum hypothesise that powers are real entities, and causation is *powers tending towards their manifestations*. In slightly more detail, Mumford and Anjum hold that 'causation happens when powers do their work' (2011: 30). Furthermore, powers do not work alone (except in exceptional cases). Most effects are the upshot of multiple powers manifesting themselves. For example, for a light bulb to burn me, the filament needs to be manifesting its power to get hot, the glass needs to be manifesting its power to propagate this heat, and my hand needs to be manifesting its liability to be burnt. Each power has a contribution to make to the coming-about-of the effect. Each power, in its own way, 'pushes' towards an effect. When many powers make their contributions, these contributions add together, and after they reach a certain threshold the effect is produced. (It is this *contribution* towards the coming-about-of some effect, not the effect that eventually comes about, which Mumford and Anjum take to be the power's manifestation. This is because Mumford and Anjum want to maintain that powers are individuated by their manifestations, so *distinct* powers cannot have the *same* manifestation, and one and the same power cannot have a different manifestation in different contexts, so Mumford and Anjum distinguish a power's manifestation from the effect of the power's manifesting itself.)

Does this make causation a relation between a power and the effect it makes a contribution towards producing? Or between the set of powers that have accumulated and the effect their accumulation has produced? Or between the power and its manifestation, i.e. the contribution it makes towards an effect? Mumford indicates that causation is 'the whole process going from power to exercise and from contribution to event' (2009: 108). He writes:

> The dispositionalist, instead of seeing causation as a matter of clearly distinguishable cause and effect, with the appropriate relation between them, sees causation as almost always complex, involving multiple powers combining to produce something together through a process. Only in the idealised laboratory conditions would we theoretically have an event produced just by one power acting alone. Instead of discrete, externally related causal relata, we have a process of interconnected powers. Given that a manifestation is a part of the essence and identity of a power, then if the power and its manifestation exists, any such causation would be an internal relation. (2009: 108–109)

Elsewhere, Mumford and Anjum state: 'We argue that causation is a single, unified, and continuous event or process *rather than a relation* between distinct and discrete events, that causes and effects are simultaneous and that causes tend towards their effects without necessitating them' (2013: 554, emphasis

added). Mumford and Anjum also describe causation as the passing or shifting of powers from one substance to another. So, when fire heats a person, the power to heat possessed by the fire is passed to the person and, when a stone breaks a window, the power to cut that the window comes to possess after this causal transaction is drawn from powers possessed by the stone. Mumford and Anjum also suggest that this process of passing around powers is more fundamental than the substances that possess the powers (2013: 555). In holding that causation is 'a single, unified, and continuous event or process rather than a relation', Mumford and Anjum seem to reject relationalism.

Misgivings about Mumford and Anjum's metaphysics have been raised by Jennifer McKitrick (2013). McKitrick objects that Mumford and Anjum's theory of causation has nothing to say about how dormant powers become active, or come to be exercised. Mumford and Anjum's view identifies causation with a continuous process of powers pushing towards an effect, but this presupposes that the powers are already being exercised—they are already making their contribution to an effect. Mumford and Anjum respond to this objection by claiming that 'when a power is not doing its work, it is not part of the causal story, so it is not something we should be trying to include' (2013: 556). They also insist that they do have something to say about how a dormant power could become active: a dormant power's becoming active could be the effect of a causal process; it could be something that resulted from the addition or removal of some other active power.

However, I think that Mumford and Anjum underestimate the seriousness of McKitrick's complaint. According to Mumford and Anjum, causal effects are achieved by the accumulation of many powers manifesting themselves and reaching a certain threshold. One may wonder how, on this picture, anything is really *produced*. On this picture, powers tend towards an effect and, once this 'tending' reaches a certain magnitude, the effect has come into being. The effect seems not to be causally produced so much as constructed, in the same way that bringing together the various parts of a statue is a way of bringing a statue into being. Mumford and Anjum deny that their view entails that causation should be thought of as a kind of ontological construction, but this denial seems inconsistent with their proposal that causation is the culmination of power-exercises adding together, as opposed to the transition from a power being dormant to a power being exercised. Although I agree with Mumford and Anjum that a primitive concept of power is needed to understand what causation is, I disagree with the ontology they propose to underpin their powers-based account of causation.

To reject relationalism, one need not go down the route that Mumford and Anjum do and deny that causation is ever a relation. To reject relationalism it is only necessary that one deny that causation is *exhaustively* constituted by a special sort of relation. One need not claim that we *never* think of causation as a relation between cause and effect. My preferred non-relationalist approach is characterised by *pluralism*. What this means is that our concept 'causation'

refers to more than one kind of thing. In my view, there *is* a distinctive sort of relation that answers to claims like '*c* is the cause of *e*' so *sometimes* what we refer to when we talk about caution is a relation. However, I think causation can also be a process *rather than* a relation, of which processes like breaking, crushing and bending are more determinate species. This view is in line with Elizabeth Anscombe's (1971) suggestion that causation is a 'highly general' determinable concept, which is an abstraction from the plethora of more specific causal concepts represented by verbs of action. I also agree with Anscombe that we come by this concept of causation when we directly perceive substances exerting causal power over other substances and associate what we see with the appropriate specific causal concept. Therefore, giving a full account of causation is not merely a matter of explaining what a relation must be like to be a causal relation. My proposal is that causation is on display not only when events make the difference to the occurrence of other events but also when substances engage in processes and thereby exercise causal powers.

Obviously, it is no good saying that causation can be a process rather than a relation without saying what a process is. I will therefore provide a metaphysical framework for my theory of causation that includes processes in its ontology. In this way, I will explain exactly what it means to say that causation is a process. I suggest that engaging in a process is analogous to instantiating a property, and that events are instances of processes.

6.2 Two concepts of causation

6.2.1 Difference-making

As mentioned, my preferred non-relationalist approach is characterised by *pluralism*, which means that our concept 'causation' refers to more than one kind of thing. 'Cause' is not an unequivocal term. In my view, there is a distinctive sort of relation that answers to claims like '*c* is the cause of *e*'. This relation can be characterised as 'difference-making'; it is the relation that obtains between an effect and that which made the difference to the effect's occurring or obtaining. The question of the exact nature of the cause–effect relation, beyond difference-making, is not something I will get into here. Pluralism leaves this question open—it says nothing specific about the nature of the cause–effect relation. Questions I shall remain neutral on include:

- Can the cause–effect relation be given a reductive analysis?
- Does the cause–effect relation hold between events or states or facts or property exemplifications or all of these?

However, there are some questions I will not stay neutral on. I want to insist that substances, including agents, cannot be relata of the difference-making

relation. The difference-making relation is distinct from the relation that obtains between a substance exercising a power and the event the substance produces in exercising that power. To give this latter relation a name, I will call it *the agency relation*—the relation that obtains between an agent and the event that agent is an agent of. To be clear, I am not denying that the agency relation cannot truthfully be called 'causal', in the sense that it has something to do with causation. What I am denying is that this relation is between 'cause' and 'effect', when 'cause' is interpreted as meaning 'that which made the difference to the effect's occurring or obtaining'. This is because substances cannot be that which made the difference to an effect's occurring or obtaining, as I shall presently show.

I also want to insist that the relation substances bear to the substances they are acting upon, a relation that I will call *the agent–patient relation*, is not a relation between cause and effect. Again, we may sometimes call the agent–patient relation 'causal' because we want to indicate that there is causation going on when the relation obtains, but the agent–patient relation is not what causation is.

Substances cannot be that which made the difference to an effect's occurring, because difference-makers must be *dated entities*. Substances, as I understand them, are entities that exist at more than one time by *enduring*. They are entities that, as it were, 'sweep through' time. They exist at multiple times (most of them, anyway) but not by having *temporal parts* located at each time. Proper temporal parts are cut out of the object along temporal dimensions but not spatial dimensions. So, temporal parts are parts that can be described as 'earlier than' or 'later than' other parts but not 'to the left of' or 'to the right of' other parts.

On the view I endorse, substances do not have temporal parts at all; they only have spatial parts. Because substances exist at more than one time by enduring, this means that substances cannot instantiate properties 'atemporally'. To take an 'atemporal' perspective on the world is to think about how the world is, while ignoring the distinction between past, present and future. It is not to think about the world as it is now, or as it was in the past, or will be in the future; it is to think about the world as it is independently of what time is 'now'. On an endurantist view of substances, substances do not instantiate (at least temporary) properties independently of what time is now. If you think about how the world is while ignoring the distinction between past, present and future, it will be impossible to say what properties substances have. It will be impossible to say, for example, whether I have blonde hair or brown hair—this is because I *had* blonde hair in the past and *now* I have brown hair.

In contrast to substances, events are paradigmatic dated entities. Events that exist at more than one time do not 'sweep through' those times; they are instead 'spread out' across those times. That is, events that exist at more than one time exist at those times by having temporal parts at those times. Importantly, events can instantiate properties atemporally—independently of what time is now. For example, the passage of time has made no difference to Roger Bannister's

record-breaking mile-run taking 3 minutes 59.4 seconds. In 1954, this event took 3 minutes 59.4 seconds, and today 3 minutes 59.4 seconds is still how long the event took. Difference-makers must be dated entities because, in looking for that which made the difference to the occurrence or obtaining of an effect, we are looking for a part of the history of the world that stands in a relation to another part of the history of the world atemporally.[24]

The fact that difference-makers need to be dated entities also shows why the agent–patient relation cannot be a relation between cause and effect. Effects also need to be dated entities, and patients, being substances, are not dated entities. Another consideration that shows why the agent–patient relation is not a relation between cause and effect is that effects must occur or obtain *after* their causes, but it is not necessary for an agent–patient relation to obtain that the patient came into existence after the agent.

When 'cause' is taken to mean 'difference-maker', substances cannot be causes, not even when they are agents. Insisting that substances cannot stand in cause–effect relations constitutes a key difference between my view and some of the agent-causationist accounts of agency discussed in the previous chapter. Unlike traditional agent-causationism, I do not claim that agents cause their own actions. Unlike the actions-as-causings view, I deny that 'an action is a causing of an event by an agent' (Alvarez & Hyman 1998: 224). When a substance is an agent they are not the cause of the event they bring about by acting. However, an important part of my non-relational account of causation is that causation is also 'the production of change by the exertion of power by a substance', a phrase I have borrowed from Thomas Reid (1788: 12–13). That is, I think that substance causation is a special kind of causation that cannot be identified with, or understood in terms of, a causal relation between events. Thus, I agree with E. J. Lowe's claim that 'a causal power, as I shall construe this term, is one whose manifestation or "exercise" consists in its bearer's acting on one or more other individual substances (or sometimes on itself) so as to bring about a certain kind of change in them (or it)' (2013: 158). How is endorsing the fundamentality of substance causation consistent with denying that substances can stand in cause–effect relations? The answer is to give substance causation a distinctive *non-relational* interpretation.

6.2.2 Substance causation

Although I think there is a distinctive sort of relation that answers to claims like '*c* is the cause of *e*', I do not think this is all there is to causation. Pluralism about causation entails that our concept 'causation' refers to more than one kind of thing. I think that causation can also be a process *rather than* a relation,

[24] Fales offers a similar explanation for why events, and not substances, are the relata of causal relations (1990: 54).

of which processes like breaking, crushing and bending are more determinate species. My proposal is that causation is on display not only when events make the difference to the occurrence of other events, but also when substances exercise causal powers, and what it is for a substance to exercise causal power is for there to be an entity, i.e. a process, in which the substance engages. As well as being a distinctive kind of relation, causation is sometimes a determinable process and substance causation is engagement in a process. Although substances are not 'causes' in the sense of difference-makers, they are *causers*, which is to say they can be, for example, movers, or breakers, or crushers, or scrapers; that is, substances can be things that engage in causal processes.

I have said that what it is for a substance to exercise causal power is for there to be an entity, i.e. a process, in which the substance engages. To fully understand how this is a non-relational interpretation of substance causation it is necessary to give an account of what processes are. The orthodox view of processes is that, if there are any differences between processes and events, they are not significant enough to warrant treating them differently in theories of causation. Many philosophers who have written extensively on causation have not paid the distinction between events and processes much or any attention.[25] Others have considered the distinction but have explicitly rejected its metaphysical significance.[26] I propose a theory of processes that denies that processes belong in the same ontological category as events.

The best way to explain what I think processes are is to start by summarising an argument put forward by Alexander Mourelatos (1978), which shows that an important subclass of verbal predications, which Mourelatos calls 'process predications', do not implicitly quantify over particulars that have (or will have) happened. After summarising Mourelatos's argument, I will suggest that what process predications implicitly quantify over are processes, which are a special kind of *universal*.

Mourelatos (1978) argues that predications can be distinguished into three semantic classes: *event, process* and *state*, the predications in each class reporting a different sort of situation or eventuality. Examples of sentences reporting events include "The sun went down" and "Roger has run a mile". Examples of process predications include "The plant is growing" and "Roger was running". And sentences that report states include "He knows Paris is in France" and "Leo loved Lauren".

---

[25] For example, Bennett, Davidson and Kim. See especially Bennett (1988), Davidson (2001) and Kim (1976).

[26] An exception may be Salmon (1984), who does take the distinctive features of processes to be important in understanding causation. However, for Salmon, 'the main difference between events and processes is that events are relatively localised in space and time, while processes have much greater temporal duration, and in many cases, much greater spatial extent' (1984: 139).

Merely considering these examples is enough to afford an intuitive grip on the differences between Mourelatos's three classes. However, Mourelatos offers a more rigorous account of the features of predicative sentences that determine which of his three classes a prediction falls into. He suggests that, when it comes to working out what sort of eventuality a sentence reports, the most illuminating feature is the grammatical aspect of the main verb (though semantic and lexical features also play a part). In Mourelatos's view, process predications typically involve verbs with progressive aspect. In English, the progressive is formed by combining the 'present' or 'ing' participle of the verb with the auxiliary verb 'be' as in "She is swimming" or "He was walking".[27] An important feature of sentences involving progressive verbs is that these sentences do not necessarily imply that the eventuality reported has or will come to an end. For example, neither "Roger was running" nor "Roger was running a mile" necessarily implies that Roger has finished, or will finish, his task. Based on the first, Roger may still be running and, on the second, Roger may still be running a mile. In the present tense, this is even clearer. "Wendy is walking" obviously does not imply that Wendy has finished walking; it implies the reverse: what Wendy is doing, walking, is still going on. Contrast this with sentences such as "Roger ran a mile", which does not have progressive aspect. It is because the progressive is often used to indicate that something is or was in progress that it is such a reliable indicator of process predications.

The fact that progressive sentences do not necessarily imply that the eventuality reported has or will come to an end allows us to draw a conclusion with metaphysical import: process predications do not implicitly quantify over particulars that have (or will have) happened. If process predications implicitly quantify over anything, what they implicitly quantify over are not particulars, or countable items. We can see this if we transform process predications into sentences that involve explicit quantification over the eventuality reported. Mourelatos calls this kind of transformation a 'nominalisation transcription' (1978: 425). For example, if we nominalise the process predication "Roger was running" we get "There was running by Roger". This nominalisation does not include an indefinite article. Similarly, the gerund "running" could not be preceded by a word like 'few' or 'many' and yield a sensible sentence. In these respects, the sentence "There was running by Roger" is akin to sentences like "There is snow on the roof" or "There is sand in the bucket", which involve mass nouns. Sentences like "There is snow on the roof" do not involve quantification over countable items; instead they involve quantification over stuff, or 'mass quantification'. The similarities between the nominalisations of process predications and quantifications over stuffs suggests that quantification involved in

[27] There is no consensus among linguists as to whether grammatical aspect is a universal feature of languages; it also appears to be encoded differently in different languages. For further discussion of grammatical aspect see De Swart (2012), Filip (2012) and Gvozdanović (2012).

"There was running by Roger" is also not quantification over countable items. As Jennifer Hornsby points out, 'the sentence "There was running by Roger" tells us that something … was going on. But it does not say of any event, nor of any particular of any other sort, that it was going on' (2012: 236). What the nominalisation of a process predication says there is (or was) is not a particular and hence not an event.

In this way, process predications stand in contrast to sentences like "Roger ran a mile". Recall that "Roger ran a mile" necessarily implies that Roger has completed the mile. When we nominalise this sentence, we get "There was a running of a mile by Roger". This nominalisation does involve quantification over particulars, and the gerund 'running' refers to a particular event. "Roger ran a mile" does say that an event (at least one) has occurred, namely Roger's running of a mile. It is for this reason that sentences like "Roger ran a mile" are classed as event predications by Mourelatos. This is also why it is plausible to argue (as Davidson (1967) does) that the sentences Mourelatos classes as event predications involve implicit quantification over events.

Mourelatos's (1978) observation that sentences reporting processes do not report the occurrence of any specific event, and involve mass quantification when they are nominalised, shows that we have a concept of a type of entity that is not particular, and hence not an event, but which exists by unfolding over time. Although one must attend to verbal predications with progressive aspect to establish that English-speakers have a concept of an entity that is not particular and which exists by unfolding over time, the presence of a process concept may be less hidden in other cultures. For example, Zhihe Wang (2013) notes that 'it is well known that Chinese thought lays great stress on process' and 'an emphasis on becoming is implicitly embodied in its understanding of Tao, the ultimate concept in Chinese tradition' (2013: 178). Wang describes Tao as 'the creative advance of the world' (2013: 178) and notes that, although Tao is translated into English as 'way' or 'path', i.e. as a noun, in Chinese the word serves as both noun and verb—it is the following of a path as much as it is a path to follow. Thus, it seems that Tao is best thought of not as analogous to Jonathan Schaffer's 'history' (2007: 83), which lacks the dynamism essential to the Tao concept, and is more similar to my concept of a highly determinable process (see Chapter 32 of the *Tao Te Ching*).

Some philosophers, inspired by Mourelatos's argument that nominalisations of process predications involve mass quantification, contend that what the process concept refers to is a kind of 'temporal stuff'. For example, Hornsby suggests that 'the relation between the stuff of the spatial world and the particulars therein is analogous to the relation between the activity [a kind of process] of the temporal world and the particulars there' (2012: 238). Thomas Crowther also maintains that '[w]hat things are doing throughout periods of time and substance stuff are constituents of the same basic ontological category; they could be thought of as temporal and spatial masses' and '[b]oth substance-stuffs and time-occupying stuffs, respectively, fill out space and time in the same way'

(2011: 17). Similarly, Helen Steward (2013) proposes that space and time have analogous ontologies. Entities that have spatial extension can be distinguished into 'stuff' and 'things'. Things are countable particulars and have spatial parts. Steward suggests there are two different types of thing: 'substances' and mere 'lumps of stuff'. Substances are entities that can survive the loss or replacement of their spatial parts; they have 'a certain distinctive form by means of which they are singled out in thought and which underwrites their relative independence from the actual parts of which they consist in any particular instant' (2013: 487). Lumps, in contrast, may be defined in such a way that 'the merest addition or subtraction [of spatial parts], however tiny, makes for a different lump' (2013: 804). In addition to things, there is also the stuff from which things are made. Examples of stuffs include snow, sand, water and clay. Stuffs are extended in space but are non-countable. The metaphysics of stuff is contentious but is not necessary to adjudicate on these questions here. What matters for now is that some philosophers have proposed that events and processes are metaphysically analogous to things and stuffs, respectively.

In what ways processes are analogous to stuffs, and to what extent they are metaphysically analogous, is an open question. Different proponents of the temporal stuff view of processes have differing opinions on how exactly to spell out the process–stuff analogy. For Hornsby, what matters is that processes are not particulars, they are distinct from events, they pervade time and they comprise events. For Steward, in contrast, processes are countable entities metaphysically analogous to substances in that they have a distinctive form that determines what intrusions, shortenings and lengthenings they could and could not have survived.

What could decide between these competing views? Mourelatos's observations about the similarities between nominalised process predications and mass nouns do not entail that processes are metaphysically similar to stuffs at all. Not all nouns that demonstrate the grammatical characteristics definitive of mass nouns obviously quantify over entities that are stuffs. For example, 'furniture' is a mass noun but (arguably) "There is some furniture in here" does not quantify over a kind of stuff; it quantifies over a collection of discrete individuals. Similarly, "There is a lot of happiness in this room" bears all the hallmarks of a mass quantification but 'happiness' is not commonly thought to refer to a stuff. Mourelatos has shown that sentences reporting processes do not report the occurrence of any specific event and involve mass quantification when they are nominalised—but this is consistent with processes being unlike stuff in every respect apart from how we typically quantify over them. Mourelatos's observations then cannot justify any specific metaphysical position on processes.

How similar you think processes and stuffs are will depend on what work you want your metaphysics of processes to do. Those who propose a temporal stuff view of processes intend this ontological scheme to help explain important concepts within philosophy of action (Crowther 2011: 6). Hornsby (2012) argues that a process ontology is key to articulating a theory of human action

that does not fall foul to the disappearing agent objection. Hornsby argues that 'the agent is given her due only when it is acknowledged that she engages in activity, where no activity is any particular' (2012: 233). She claims that 'one needs to think of a person's raising her arm as a type of causal activity in which she engages' (2012: 234). Hornsby's view is that to properly understand agency we need to think of the causality of action as something other than a causal relation between mental event and action—a proposal I agree with. Hornsby further suggests that construing an agent's causing something (for example, an agent's causing her arm to go up) is an activity or process—something that is metaphysically distinct from an event—allows us to think of the causality of action as something that essentially involves the agent herself. The causality of action is thus thought of as a unique sort of entity: an activity or process. Once we acknowledge this, we are no longer at risk of failing to include the agent in an account of the causality of her action.

I agree with Hornsby's explanation of what work a metaphysics of processes is supposed to do. For me, the justification for adopting a process ontology is that doing so helps articulate a non-relational theory of causation. More specifically, the point of proposing a process ontology is to help explain what substance causation is, which will in turn allow us to put together a theory of agency that recognises the essential role of the agent in the causality of action. The process ontology that I think fulfils this mandate most effectively is, in fact, *not* one that takes processes to be 'temporal stuffs' that pervade intervals of time and compose events in the same way that spatial stuffs pervade volumes of space and compose things (see White (2020) for an argument against the temporal stuff view of processes).

I submit that what the process concept refers to is a special kind of *universal*. Processes are universals, so running, singing, respiring and melting are single repeatable entities; when Usain Bolt is running, the very same entity is present, or going on, as when Roger Bannister was running. This is not to say that processes are properties, which are also thought to be universals by some philosophers (including Armstrong 1978a; Armstrong 1978b; Armstrong 1989). The distinction between processes and properties can be drawn in the following way: properties concern the static nature of things—they are 'ways for things to be'—whereas processes are dynamic, that is, they are connected with how a thing is changing over time (White 2020). My proposal is that processes are ways for a substance to be changing, to be resisting change, or to be effecting change. The last subgroup of processes is particularly important and these types of processes are what I believe are picked out by the concept 'activity'.

My theory of processes is outlined in White (2020). There I proposed that *process, event* and *substance* are three distinct ontological categories. I proposed that processes are engaged in by substances. According to this ontological scheme, 'A process P exists, or rather goes on, only when, and for as long as, a substance engages in P', a principle I called 'the engagement principle' (White 2020: 118). This means that processes depend for their existence on substances engaging in

them. I also suggested that, when a substance engages in a process, this unity of substance and process can be called a *dynamic* state of affairs (White 2020: 119). Dynamic states of affairs bear a similarity to the states of affairs that feature in Armstrong's account of properties. Armstrong claims that, when a substance instantiates a property, which he takes to be a universal, the unity of substance and universal is a 'state of affairs' (Armstrong 1989: 88). The difference between Armstrong's *static* state of affairs and my *dynamic* state of affairs is the relationship that dynamic states of affairs bear to time.

Static states of affairs persist (continue to exist) by enduring over time. This means that they do not have temporal parts. Instead, static states of affairs exist complete at the instant at which they first obtain, and then continue to exist by continuing to obtain. For example, the state of affairs of this rose's being red exists complete at the instant at which it is true that the rose is red—no part of the state of affairs exists at any other time. Dynamic states of affairs, on the other hand, cannot exist 'complete' at a single moment. Because processes are concerned with how a substance is changing, resisting change, or effecting change, engaging in a process presupposes the passage of time: nothing can be going on for only an instant (although, of course, it can be true *at* an instant that something is going on). This means that dynamic states of affairs cannot obtain for only an instant. If a dynamic state of affairs obtains, then necessarily time has passed or will pass. For example, no-one can be running for only an instant. To run, one must make the right sort of leg movements—one needs to raise one leg, lift off from the other, land on the first, transfer weight, and so on—it is impossible to accomplish this in an instant. If someone made the first movement of running, i.e. raised one leg but got no further than this, then we would deny that that person was ever running. If someone is running, this conceptually implies something about the past or the future. Dynamic states of affairs are, in a sense, stretched out in time. Their obtaining is necessarily dependent on the passage of time. This suggests that dynamic states of affairs do not persist by enduring.

The alternative to persisting by enduring is typically assumed to be to persist by perduring. This means to exist at more than one time by having temporal parts that exist at more than one time. An event persists through time by perduring—it has earlier stages at earlier times and later stages at later times. It is spread out through time. Is this how dynamic states of affairs persist over time? This might seem like an obvious answer, especially given that dynamic states of affairs seem to be 'stretched out in time', but in fact I think it is incorrect. I do not think dynamic states of affairs persist by perduring. Things that persist by having temporal parts are things that happen. Part of what it means to say an entity is an occurrence, something that happens, is that it is temporarily extended and has temporal parts. As Steward (2013) argues, denying this leaves us with no clear way of drawing the distinction between things that occur and things that exist at more than one time by enduring. However, dynamic states of affairs do not happen; they *obtain* and it seems to me that obtaining and happening

are mutually exclusive modes of existence. The relationship that dynamic states of affairs bear to time is thus not straightforward. It is neither enduring nor perduring but something in between. It is very difficult to articulate what this could be. Dynamic states of affairs seem to be stretched out in time but they do not have temporal parts. Indeed, as states of affairs, it seems incoherent to talk of dynamic states of affairs as having parts at all. Dynamic states of affairs have *components*—namely, a substance and a process (a universal)—but not *parts*.

In White (2020) I also proposed that events are *instances* of processes where instancing is analogous to the relationship between kinds, like doghood, and individual substances, like an individual dog (see Lowe's (2005) four-category ontology) and to the relationship between a pattern, like a wallpaper pattern, and a physical realisation of this pattern, for example a piece of wallpaper (as in Galton's (2018) theory of processes as 'temporal patterns'). Events come into existence when a substance engages in a process and then completes or stops the process. For example, when a tank crushes a car, the tank engages in the process of crushing for a certain length of time (e.g. until the car is crushed) and once that process is complete or stopped a crushing event can be said to have happened.

It is dynamic states of affairs that are reported by Mourelatos's process predications. It is also dynamic states of affairs that, I propose, are referred to by expressions such as 'the agent caused her arm to rise'. The infinitival phrases that are commonly used to describe exercises of agency refer to dynamic states of affairs. Phrases like 'the agent caused her arm to rise' should not be taken to mean that a relation of causation comes to obtain between the agent and an arm-rising event. Thus, unlike the accounts of substance causation outlined in Chapter 5, what it is for an agent to be causing something is not for that agent to cause an event to happen.

I have said that, as well as being a distinctive kind of relation, causation is sometimes a determinable process that substances engage in. What it is for a substance to exercise causal power is for there to be an entity, i.e. a process, in which the substance engages. However, not all processes are examples of causation. If any process is a determination of causation, then it is causal *intrinsically*, just as if a colour is a determination of red (as scarlet is), then that colour is red intrinsically. As to *which* processes are determinations of causation and which aren't, my answer is that the distinction is not absolute, and can be difficult to determine.

I have said that processes are ways for substances to be changing, to be effecting change or to be resisting change. This means that some processes are active, i.e. those that are ways for substances to be effecting change, and some processes are passive, i.e. those that are ways for substances to undergo change (resisting change, I think, can be both active and passive). Only those processes that are (to some degree) ways for substances to be effecting change are species of causation. This way of distinguishing between processes that are causal and those that are not makes use of the distinction between active and passive powers.

An active power is a power to wreak change. Activity is the exercise of an active power. A passive power, or a liability, is a power to undergo or suffer change. Passivity is the manifestation of a passive power. Active powers are powers to change, and passive powers are powers to be changed. Substances that exercise active powers are agents, and substances that manifest passive powers are patients. The difference between agent and patient is not a difference between two different kinds of substance; it is rather a difference between two different roles substances can adopt (Hyman 2015: 35). This is demonstrated by the fact that one and the same substance can be an agent at one time, and a patient at another time—for example, when I push you, I am the agent, when you push me back, I am the patient. It is also possible for one and the same substance to be both agent and patient at the same time—for example, as Hyman notes, a victim of suicide is both agent and patient.

The active–passive distinction is thrown into doubt when we consider the fact that in many cases when an intuitively active power is manifested the manifestation of this power involves the possessor of the power suffering change as well as producing it. For example, when salt is dissolved in water, we may intuitively class the power of the water to dissolve the salt as active: the water is producing change in the salt. However, the water is also changed by the dissolution process, and necessarily so—if the water were not liable to become uniformly salty when salt was added to it, then it wouldn't be possible to dissolve salt in water. So, it seems that the intuitively active power of water to dissolve salt is *also* passive. It seems like the distinction between the exercise of active power and the manifestation of passive power, and hence the distinction between activity and passivity, is spurious. At best, the distinction is a matter of there being two alternative ways to describe the very same sort of eventuality.

The solution to this problem is, I think, to reject the idea that for a substance to exercise an active power the substance must, in exercising this active power, be 'purely active', that is, suffer no change at all. Similarly, it is not the case that a substance exercising a passive power needs to be 'purely passive'. Erasmus Mayr suggests that 'the distinction between active and passive powers is one of degree, with all powers situated on a more or less continuous spectrum of more or less active and passive powers' (2011: 204). What this means is that some powers are such that when they are exercised the substance in possession of the power produces much more change than it undergoes. For example, when I squash a grape, the grape is drastically changed, whereas I remain much the same. Other powers are such that when they are exercised the substance in possession of the power undergoes as much change as it produces—as in the case of the water dissolving the salt. The power of the water to dissolve salt is, as it were, less active than my power to squash a grape.

The danger with this solution is that it means that the distinction between activity and passivity is not absolute. It is therefore more accurate to say that some processes are more active than others, and some are more passive than others, but (probably) no process is completely active, and no process is

completely passive. For example, the process of crushing something is mostly active: in crushing something, a substance is effecting more change than it is undergoing. The process of dying, on the other hand, is mostly passive: in dying, a substance is undergoing more change than it is effecting. And many processes involve ostensibly equal degrees of activity and passivity. For example, processes by which we move ourselves about, like walking, and running, seem to involve a mix of activity and passivity: when we move ourselves about, we effect change on ourselves, so we are both agent and patient with respect to those changes. Processes that result in no overall change, like thermoregulation or keeping still, also seem to involve elements of activity and passivity. When one stands still, for example, one must exert some degree of force in opposition to the forces that would cause one to fall to the ground (e.g. gravity), but not so much force that one ends up moving. Thus, standing still seems to involve a roughly equal mix of activity and passivity.

The mostly active processes I will call *activities*. What it is for a substance to be causing something is for there to be an *activity* that the substance is engaging in. A substance engaging in an activity is an agent, and the event that results once the substance has completed the activity it has been engaging in is an action. For example, when I crush a grape between my fingers, I engage in the activity of crushing. When I complete that activity (when the grape is crushed), a crushing action can be said to have happened. Actions are thus events of a special kind: they are events that are instances of activities, and as engaging in an activity is what it is for an agent to be causing something, actions can also be said to be instances of substance causation. Importantly, the agent does not stand in a cause–effect relation to the event that comes into existence after she completes an activity. Agents are not causally related to their actions. Individual actions are events that come into existence when an agent engages in an activity and then completes that activity. So understood, actions are 'produced by' or 'brought into being by' agents, but the sense of production here is a kind of ontological construction. Actions depend for their existence on agents engaging in activities and completing them, so actions come into existence because of agents engaging in activities—but this 'because' indicates ontological rather than causal dependence.

Another issue with the active–passive distinction is that it is less than fully objective. Whether what a substance is doing is activity or passivity is relative to the degree of change it is wreaking and/or undergoing and assessing how much change a substance is wreaking and/or undergoing may not be a fully objective matter. How much change one thinks the water undergoes when salt dissolves into it may depend on one's views about the nature of water. If the distinction between activity and passivity is partly a subjective matter, and this distinction is key to distinguishing processes that are determinations of causation from processes that are not, then it seems that what is and is not causation is itself partially a subjective matter. I think that this reasoning is sound, so I accept that what is and is not causation is partially a subjective matter.

However, I do not consider this to be problematic. This is because, while it may be true that how we classify the processes being engaged in by substances is partly dependent on our own perspective, the *existence* of dynamic states of affairs, i.e. substances engaging in processes, is not mind-dependent.

The notion of 'effecting change' is clearly a causal notion, hence my account of substance causation cannot be reductive. However, I deny that my account is circular (i.e. analyses causation in terms of causation). This is because we are acquainted with the determinate forms of causation (like breaking and crushing) via direct observation, and, to borrow an argumentative strategy from Peter Menzies and Huw Price, 'this common and commonplace experience … licences what amounts to an ostensive definition' of effecting change (1993: 194). We directly observe the determinate forms of causation, which allows us to point to an example of a substance effecting change and say '*that* is what effecting change is'. In this way, it is not necessary for one to already understand what causation is before one can know what 'effecting change' is.

Rom Harré and Edward Madden (1975) also argue that we directly perceive processes in which causal powers are manifested. They argue that David Hume's denial that we directly perceive powers being exercised is based on the false assumption that our perceptual experience is primarily atomistic. Hume assumes that what we directly experience are 'punctiform', 'atomistic' sensations. Once this assumption is made, it follows that it is impossible that a single impression could be the experiential origin of our idea of causal power, and hence some story must be told about how the idea of causal power arises from multiple impressions. However, why assume that our singular impressions are all and only 'punctiform', 'atomistic' sensations? Why assume that we directly perceive the leaf as green and, later, the leaf as brown, but that we do not perceive that leaf *changing* from green to brown? Anscombe objects to Hume's idea that we cannot observe causality in the individual case by pointing out that 'someone who says this is just not going to count anything as "observation of causality"' (1971: 8). Anscombe is, I think, making a very similar point to Harré and Madden. If one assumes from the outset that perceptual experience is primarily atomistic, then of course it will turn out that 'all we find' are impressions of events that 'seem entirely loose and separate' (Hume 1975: 74), but that's because 'the arguer has excluded from his idea of "finding" the sort of thing he says we don't "find"' (Anscombe 1971: 8).

6.3 Objections to pluralism

I believe pluralism is the best way to do justice to the diversity of our causal thinking. When it comes to explaining why the relation between the collision with the iceberg and the sinking of the ship, or the relation between the fluttering of the flag and the bull's charging, are instances of causation, appeals to powers and their exercise may not provide the answer. (Appeals to powers

and their exercise may explain why such relations exist, without explaining what the relations actually are.) On my view, there is no demand to provide a semantics for all causal discourse in terms of powers. I can allow that the conceptual scheme that relates the concepts *power, substance* and *process* may not (and, I suspect, cannot) be sufficient to clarify the content of *all* our causal claims.

The idea that we have more than one way of thinking about causation is not such a novel idea. Brian Skyrms has suggested that, rather than being a single concept, causation is an 'amiable confused jumble' of concepts (1984: 254). My view honours this suggestion: on my view the concept 'causation' covers an ontologically diverse 'jumble', including a distinctive cause–effect relation and a determinable process, which is in turn associated with two distinctive sorts of relation, the agency relation and the agent–patient relation.

My view is perhaps most similar to a position put forward by Richard Taylor (1966). In his introduction to *Action and Purpose*, Taylor distinguishes between two meanings that have been attached to the words 'cause' and 'causation'. On the one hand, there is a notion of causation that is tied up with notions of power, which was once regarded as a 'basic' concept 'more obvious and more clear than any concepts by means of which one might try to describe or define it' (1966: 16). On the other hand, there is the notion of causation as a 'complex relationship between changes or events, analysable in terms of other familiar relations such as constant conjunction and not, in any case, one that can be understood only in terms of some further primitive notion of active power, or the power to make things happen' (1966: 16).

One potential objection to my view is that the idea that we think of causation in two distinct ways—as a process and separately as a cause–effect relation—is inconsistent with the fact that we use just one word, 'causation', to cover the worldly phenomenon. As Randolph Clarke presents the objection:

> To say that entities of both these categories [substance and event] can be cause is to say that causation can work in two dramatically different ways. Causation would then be a radically disunified phenomenon. It may be claimed, with some plausibility, that this cannot be so. (2003: 208)

I think this objection can be dealt with by acknowledging that, even though we think of causation in two different ways, our two causation concepts are not entirely disconnected from each other. One way to spell out this claim is to offer a plausible story of how one of the two causal concepts may have grown out of the other. The story I find the most plausible runs as follows. As noted above, if substances possess and exercise causal powers, then substances with certain powers must behave in certain ways when the conditions for the manifestation of the power arise, provided there is nothing interfering. In other words, when a power is properly triggered, it will manifest itself in 'canonical ways', as Nancy Cartwright puts it (2009: 144). The exercise of powers will therefore be the source of regular and stable relations between trigger events and manifestation

events. We can use knowledge of these relations to change how powerful substances behave. For example, if one knows that being near flowers triggers an allergic reaction, then one can prevent the allergic reaction by avoiding flowers; similarly, if one knows that a release of luteinising hormone by the pituitary gland triggers ovulation, then one can prevent ovulation by preventing the release of luteinising hormone. From this we get the idea that events, particularly (but not exclusively) trigger events, can be *devices* for manipulating later events and can *produce* later events. However, this is a metaphor: events are not literally devices, and cannot literally produce events because they are not the right sort of thing to be devices or produce events—only substances can literally play these roles. This is because producing an event is a process. A trigger event cannot produce a manifestation event because the manifestation event occurs after the trigger event is over and done with—the trigger event is in the past when the manifestation event begins to occur, hence the trigger event is not around at the right time to produce it. Only something that endures for the occurrence of an event can produce it. However, even though talk of events as devices or producers is a metaphor, this does not mean there aren't conditions under which use of this metaphor is correct and conditions under which use of this metaphor is incorrect, just as the fact that feelings can only metaphorically be hurt does not mean it is never incorrect to say my feelings have been hurt. This metaphor is thus the source of the idea that there is a special sort of relation between events, which is causation.

Another objection to pluralism is that the two ways of thinking about causation I have proposed are not both needed. One might think that the concept of difference-making is sufficient to fully capture the concept of causation, or alternatively that the concept of substance causation is sufficient to fully capture the concept of causation.

The reason I do not think that difference-making on its own is sufficient to fully capture the concept of causation is because, as Steward puts it, 'an important aspect of our conception of causation seems to involve the idea that causes do things' (2011: 152). Here we seem to have a platitude included within our concept of causation that ascribes to causes the power to do things. In agreement with Lowe (2013), I do not think that events are the sort of entity that possess causal powers. Lowe's definition of a causal power, which I think is correct, is a power 'whose manifestation or "exercise" consists in its bearer's acting on one or more other individual substances (or sometimes on itself) so as to bring about a certain kind of change in them (or it)' (2013: 157). Given that this is what a causal power is, only entities that can act on substances or themselves could possess causal powers and, as Lowe correctly points out, 'events and properties cannot literally act: only substances can do that' (2013: 158).

Lowe argues that, 'fundamentally speaking, all causation is substance causation, because only substances strictly and literally possess causal powers' (2013: 157). Lowe suggests that event causation is unnecessary:

> [W]e might say, for instance, that the explosion of the stick of dynamite caused the collapse of the building. But really, in my view, this is just an elaborate way of saying that the stick of dynamite, by exploding, caused the building to collapse. It is the dynamite that literally possesses the destructive power, not the explosion. (2013: 158)

Lowe seems to think that the concept of substance causation sufficient to fully capture the concept of causation, and therefore that the difference-making understanding of causation is unnecessary. My reply to Lowe is first that it would be a mistake to infer from the fact that only substances strictly and literally possess causal powers that substance causation is the only kind of causation that exists. For events to be causes, they need to be that which made the difference to the occurrence of an effect—they do not need to strictly and literally possess causal powers. Lowe's characterisation of causation as 'a kind of action—a bringing about of change' is a good description of substance causation, but not of difference-making. Second, although I admit that it does seem frivolous to hold both that the dynamite caused the collapse and that the explosion of the dynamite caused the collapse, in fact there isn't any kind of competition between the dynamite's causal efficacy and the explosion's causal efficacy, because substances and events take part in very different kinds of causation. The dynamite brought the collapse into being by engaging in the process of destroying the building by exploding. The explosion is the event that stands in a difference-making relation to the collapse.

6.4 Objections to substance causation

In the previous section I considered objections to pluralism, the idea that there is more than one kind of causation and that the term 'causation' does not have a single meaning. In this section I will consider objections to my non-relational understanding of substance causation.

Insofar as my view grants that causation can be an exercise of causal power, my view has a lot in common with powers-based theories of causation such as that proposed by Mumford and Anjum (2011) and by Lowe (2013). Like these writers, I also maintain that *power* is a primitive concept, i.e. one that cannot be analysed in other terms. So, one cannot say, in other terms, what is meant by 'can' in statements of what a thing can do. As other powers-based theories of causation maintain, I think that facts about what powers things have, or what things can do, cannot be analysed as claims about what events regularly follow on from what others. Instead, causation is something in nature that constrains the ways in which events can unfold, and which therefore *grounds* regularity. In other words, worldly events unfold in a regular way *because* causation exists. Causation is the exercise of power and worldly events unfold in a regular way because what can occur is limited by what powers entities possess: an entity

with certain powers must behave in certain ways when the conditions for the manifestation of the power arise, provided there is nothing interfering with the entity and thereby blocking the manifestation.

However, unlike Mumford and Anjum, I do not think that powers are *entities*. Powers do not exist in concrete reality; they are not, to borrow a phrase from Lowe (2005: 35), 'elements of being'. This idea has been expressed by Anthony Kenny, who states that 'a power must not be thought of as a thing in its own right' (1975: 10) and by Gilbert Ryle, who states that:

> Potentialities, it can be truistically said, are nothing actual. The world does not contain, over and above what exists and happens, some other things which are mere would-be things and could-be happenings. (1949: 119)

In agreement with Ryle, I deny that ascriptions of powers to things report 'limbo facts' or strange nearly-properties. However, as Ryle puts it, 'the truth that sentences containing words like "might", "could" and "would … if" do not report limbo facts does not entail that such sentences have not got proper jobs of their own to perform' (1949: 120). The concept *power*, it seems to me, is best thought of as a way of thinking about how substances are connected to the processes they engage in, not just currently but possibly in the future and in circumstances that may never come to pass. As Ryle contends, the job of ascriptions of power is to allow us to make inferences about what substances can, will and would do.

Because my view has a lot in common with powers-based theories of causation it risks falling foul of the same objections. For example, Jonathan Schaffer (2007) objects to the idea that worldly events unfold in a regular way, because what can occur is limited by what powers entities possess. According to Schaffer, such a view places implausible limits on what can be. Schaffer regards the view that 'anything can coexist with anything else, at least provided they occupy distinct spatiotemporal positions' (Lewis 1986: 87), as a 'plausible principle about what is possible' (2007: 85). The idea that what can happen is limited by what powers things possess entails 'implausible limitations on recombination'; for example: 'if *c* is accorded the basic property of *causing e*, then the intuitive possibility of *c* without *e* is lost' (Schaffer 2007: 85). However, I do not think facts about what powers entities possess place implausible limits on what can be. To borrow an example from Harré and Madden (1975), if fire has the power to burn a person, and the conditions for the manifestation of this power are met, e.g. a person has stepped into the fire, what this means is that, *unless something interferes*, the person will get burnt. Is that an implausible limitation on what can be? I do not think so. And, as for Schaffer's own example, if some substance is engaged in the process of causing *e*, this does not imply that the possibility of the substance existing without *e* occurring is lost. While the substance is engaged in the process whose completion eventually constitutes occurrence

of *e, e* has not yet been caused, and may never be caused: something could interrupt the process, and *e* may never come to be. Interventions are nearly always possible, so the manifestation can be blocked by an intervention.[28] So, this objection of Schaffer's fails.

Another common objection to powers-based theories of causation is that such theories are ontologically profligate. That is, they posit the existence of fundamental sorts of entity, or make use of unanalysable concepts, to no explanatory advantage. Schaffer suggests that theories like mine involve a 'terrible metaphysical price for a relatively flimsy intuition' (2007: 89). It is important to be clear on what the metaphysical price of my theory is.

The price involves an ideological and an ontological component. The ideological element is the primitive power concept that I think we need to understand causation: I am maintaining that there are facts about what substances can do, which we can discover, where the notion of 'can' here cannot be analysed in other terms. The ontological element is the process ontology I am proposing: I am positing the existence of processes; as well as the history of events, there is also the bringing-about of those events. And what do we get for this price? The motivation for proposing an alternative to the relational approach to causation is to enable us to fully break out of the physicalist triad. I will leave it to the reader to judge whether this constitutes a 'terrible metaphysical price for a relatively flimsy intuition'.

References

Alvarez, M and Hyman, J 1998 Agents and their actions. *Philosophy*, 73(284): 219–245. DOI: https://doi.org/10.1017/s0031819198000199

Anscombe, G E M 1971 *Causality and determination: An inaugural lecture*. Cambridge: Cambridge University Press.

Armstrong, D M 1978a *Universals and scientific realism*. New York: Cambridge University Press.

Armstrong, D M 1978b *A theory of universals. Universals and scientific realism volume II*. New York: Cambridge University Press.

Armstrong, D M 1989 *Universals: An opinionated introduction*. Boulder, CO: Westview Press.

Bennett, J 1988 *Events and their names*. Indianapolis, IN: Hackett.

Cartwright, N 2009 Causal laws, policy predictions, and the need for genuine powers. In: Handfield, T *Dispositions and causes*. Oxford: Oxford University Press.

Clarke, R 2003 *Libertarian accounts of free will*. New York: Oxford University Press.

Crowther, T 2011 The matter of events. *Review of Metaphysics*, 65(1): 3–39.

[28] Mumford and Anjum (2010) make a similar argument.

Davidson, D 1967 Causal relations. *Journal of Philosophy*, 64(21): 691–703. Reprinted in Davidson 2001 pp. 149–162.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

De Swart, H 2012 Verbal aspect. In: Binnick, R I *The Oxford handbook of tense and aspect*. New York: Oxford University Press. pp. 752–781.

Fales, E 1990 *Causation and universals*. London: Routledge.

Filip, H 2012 Lexical aspect. In: Binnick, R I *The Oxford handbook to tense and aspect*. New York: Oxford University Press. pp. 721–752.

Galton, A 2018 Processes as patterns of occurrence. In: Stout, R *Process, action, and experience*. Oxford: Oxford University Press.

Gvozdanović, J 2012 Perfect and imperfect aspect. In: Binnick, R I *The Oxford handbook to tense and aspect*. New York: Oxford University Press. pp. 781–803.

Harré, R and Madden, E H 1975 *Causal powers*. Oxford: Blackwell.

Hornsby, J 2012 Actions and activity. *Philosophical Issues*, 22(1): 233–245. DOI: https://doi.org/10.1111/j.1533-6077.2012.00227.x

Hume, D 1975 *Enquiries concerning human understanding and concerning the principles of morals*. 3rd ed. Oxford: Clarendon Press.

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Kenny, A 1975 *Will, freedom, and power*. Oxford: Blackwell.

Kim, J 1976 Events as property exemplifications. In: Brand, M and Walton, D *Action theory*. Dordrecht: D. Reidel. pp. 310–326.

Lewis, D K 1986 *Philosophical papers*. New York: Oxford University Press.

Lowe, E J 2005 *The four-category ontology: A metaphysical foundation for natural science*. Oxford: Clarendon Press.

Lowe, E J 2013 Substance causation, powers and Humean agency. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 153–173.

Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.

McKitrick, J 2013 Getting Causes from Powers by Stephen Mumford and Rani Lill Anjum. *Analysis*, 73(2): 402–404. DOI: https://doi.org/10.1093/analys/ant016

Menzies, P and Price, H 1993 Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44(2): 187–203. DOI: https://doi.org/10.1093/bjps/44.2.187

Mourelatos, A 1978 Events, processes and states. *Linguistics and Philosophy*, 2(3): 415–434. DOI: https://doi.org/10.1007/bf00149015

Mumford, S 2009 Passing powers around. *The Monist*, 92(1): 94–111. DOI: https://doi.org/10.5840/monist20099215

Mumford, S and Anjum, R L 2010 A powerful theory of causation. In: Marmodoro, A, *The metaphysics of powers: their grounding and their manifestations*. New York: Routledge. pp. 143–159.

Mumford, S and Anjum, R L 2011 *Getting causes from powers*. New York: Oxford University Press.

Mumford, S and Anjum, R L 2013 Causes as powers: Book symposium on Stephen Mumford and Rani Lill Anjum: Getting Causes from Powers. *Metascience*, 22(3): 545–559. DOI: https://doi.org/10.1007/s11016-013-9783-5

Reid, T 1788 *Essays on the active powers of man*. Edinburgh: John Bell, Parliament-Square, and London: G G J & J Robinson.

Ryle, G 1949 *The concept of mind*. London: Hutchinson's University Library.

Salmon, W C 1984 *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Schaffer, J 2007 Causation and laws of nature: Reductionism. In: Sider, T, Hawthorn, J and Zimmerman, D W *Contemporary debates in metaphysics*. Malden, MA: Blackwell, pp. 82–107.

Skyrms, B 1984 EPR: Lessons for metaphysics. *Midwest Studies in Philosophy*, 9(1): 245–255. DOI: https://doi.org/10.1111/j.1475-4975.1984.tb00062.x

Steward, H 2011 Perception and the ontology of causation. In: Roessler, J, Lerman H and Eilan, N *Perception, causation, and objectivity*. Oxford: Oxford University Press.

Steward, H 2013 Processes, continuants, and individuals. *Mind*, 122(487): 781–812. DOI: https://doi.org/10.1093/mind/fzt080

Taylor, R 1966 *Action and purpose*. Englewood Cliffs, NJ: Prentice-Hall.

Wang, Z 2013 *Process and pluralism: Chinese thought on the harmony of diversity*. Berlin: Walter de Gruyter.

White, A 2020 Processes and the philosophy of action. *Philosophical Explorations*, 23(2): 112–129. DOI: https://doi.org/10.1080/13869795.2020.1753801

Woodward, J 2003 *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

CHAPTER 7

Causal Explanations

In the previous chapter I outlined a non-relational metaphysics of causation. According to this theory, causation is not always and everywhere a relation but can be a process that substances engage in. I presented a novel metaphysical framework, which includes processes, conceived of as universals, in its ontology. This metaphysical framework gives content to the claim that causation can be something substances engage in, rather than merely an external relation holding between events (or any other particulars). In the following chapters I will argue that this alternative way of thinking about causation, and the ontology that permits it, allows us to put together a new theory of intentional action and the mental causation associated with it. The ultimate aim of this theory will be to show that it is possible to reject the relational understanding of mental causation: as-a-cause is not how we should understand the place of mentality in intentional action. Intentional action does not entail the existence of causal relations between mental items and physical events.

It is commonly held that we can achieve an adequate account of what it is to act intentionally by examining the distinctive sort of explanation with which intentional actions are associated. Part of what makes intentional action distinctive is that we can explain why someone acted intentionally by giving their reason for acting as they did. Such explanations are called 'rationalising explanations'. Therefore, the path to concluding that intentional action does not involve causal relations between mental items and physical events involves challenging Davidson's claim that 'the primary reason for an action is its cause' (1963/2001: 4). We saw in Chapter 2 that there were two parts to the conclusion of Davidson's (1963) argument concerning rationalising explanations. First, rationalising explanations give causal information. Second, rationalising explanations are true if and only if the belief or desire that explains the action stands in a causal relation to the action explained. We also saw that construing rationalising explanations as explanations that

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 141–154. York: White Rose University Press. DOI: https://doi.org/10.22599/White.h.

posit an entity that is causally related to the action explained encourages us to accept an ontology that includes mental items that stand in causal relations to human actions. If we also assume that actions are physical events, for example bodily movements, then Davidson's position entails the relational understanding of mental causation.

Davidson's view that states of desiring and states of believing are causes of the actions they explain has been challenged before. Non-causalists reject the idea that beliefs and desires stand to actions as causes to effects. On this view, concepts like *belief, desire* and *intention* do not refer to items that can stand in causal relations to actions or physical events, so when such concepts are employed to explain why an agent acted they do not designate inner causes of the action they explain. However, non-causalists reach this conclusion by arguing that rationalising explanations of intentional actions are not causal explanations at all. In other words, non-causalists reject the second of Davidson's conclusions by rejecting the first.

Even though I agree with non-causalists that concepts like *belief, desire* and *intention* do not signify or denote inner causes of the actions they explain, I believe that rationalising explanations of intentional actions do give causal information. Fortunately, this kind of view, which is intermediary between Davidsonians and non-causalists, is made possible if one rejects the relational approach to causation. In this chapter, I show that it is not necessary for an explanation to be causal that its explanandum designate an effect and its explanans designate an item that is the cause of that effect. My non-relational theory of causation implies that facts about causal relations between events are not the only causal facts that causal explanations could answer to. I suggest that some causal explanations are made true by the non-relational aspect of causal reality, that is, by facts about substances engaging in processes. In Chapter 8, I will argue that explanations of intentional action that cite the agent's reasons for acting are the kind of causal explanation that are not made true by causally related events and explain why this position is preferable to the non-causalist position.

7.1 Four counterexamples to the Davidsonian view

Davidsonians and non-causalists alike assume that causal explanations are precisely those explanations whose explanandum designates an effect and whose explanans designates an item that is the cause of that effect. William Child describes the Davidsonian view as follows:

> The general idea, then, is that the truth (or acceptability) of a causal explanation rests on the presence of appropriate relations of causation. And a natural thought would be to put the point in the following way: a causal explanation is one whose explanatory power depends on

> the assumption that there are events mentioned, or pointed to, in the explanans and explanandum sentences, between which the natural relation of causation obtains; and whose truth (or acceptability) requires that the relation does indeed obtain. (1994: 102)

This view assumes that a causal *explanation* is the statement of a non-natural, intentional relationship that holds between true propositions. The causal *relation*, in contrast, is a natural, extensional relation that 'holds in the natural world between *particular events or circumstances*, just as the relation of temporal succession does or that of spatial proximity' (Strawson 1985: 115, emphasis added). This theory does not demand that the events, whose causal connectedness grounds the truth of a causal explanation, should be *explicitly* referred to or mentioned by the sentences that form the explanandum and explanans of the causal explanation, or that the explanandum and explanans sentences can be transformed into sentences that involve explicit quantification over events.[29] As Child notes, 'the fact that, in some (or even most) cases, reference to causally related events is concealed is compatible with the idea that the truth of an explanation depends on the presence of appropriate relations to causality between particular events' (1994: 102). However, the assumption is that it is necessary for an explanation to be causal that its explanandum designate an effect and its explanans designate an item that is the cause of that effect.

I will outline four kinds of counterexample to the Davidsonian view of what makes explanations causal. Then I will show why these causal explanations are best understood as being made true by the non-relational aspect of causal reality, that is, by facts about substances engaging in processes.

7.1.1 Negative causal explanations

The first counterexamples to the Davidsonian view are negative causal explanations, i.e. causal explanations where either the explanans or the explanandum, or both, is a fact about an event failing to occur.

(a) Don did not die because his rope did not break. (Child 1994: 106)
(b) The water swept away the fish because the sluice gate did not shut.
(c) The policeman was not hurt because the bullet got stuck in his Kevlar vest.

[29] The process of transforming a sentence like "Roger ran a mile" into a sentence that explicitly quantifies over an event ("Roger's running of a mile") is a process Mourelatos calls 'nominalisation transcription'. Nominalisation transcription is discussed in Section 6.2.2.

On the Davidsonian view, these explanations are causal explanations if and only if they are made true by a causally related pair of events. But in (a) it seems like no events are mentioned or pointed to by the explanation, in (b) the explanans clause does not seem to mention an event, and in (c) the explanandum clause does not seem to mention an event. One could respond by positing 'negative events'. This allows one to argue that in fact the explanans clauses and the explanandum clauses of (a)–(c) do all explicitly mention events whose causal connections serve as truth-makers for the explanations. However, as I argued in Section 4.1.2, on any theory that takes seriously the idea that events are happenings, something's not-happening cannot be an event.[30]

A more plausible response to negative causal explanations is suggested by Child. Child suggests that the Davidsonian could potentially accommodate negative causal explanations within his account of causal explanations by allowing the relation between a causal explanation and the causally related events that make the explanation true to be opaque (1994: 106). The Davidsonian position is safe if the truth of negative causal explanations depends on there being causal relations between events; it is not necessary that the negative causal explanation itself mention the causally related pairs of events that make it true. The idea would be that "Don did not die because his rope did not break" succeeds as an explanation only because rope-breakings are causally related to deaths when they occur in circumstances similar to Don's—the explanation depends for its truth on causal relations between rope-breakings and deaths. Another way of putting this point is to say that negative causal explanations are true when they are backed by a causal law—i.e. a generalisation that says that events of one type always (or usually) cause events of another type to occur.[31] This response is structurally similar to Clarke's (2014) account of how the intentionality of refrainment still depends on mental states causing actions even though refrainments themselves are the absence of an action and therefore not the sort of thing that can be caused.

There is nothing wrong with the idea that the relation between an explanation and what makes the explanation true can be opaque. As Kevin Mulligan, Peter Simons and Barry Smith put it, it is 'perfectly normal for us to know *that* a sentence is true, and yet not know completely *what* makes it true' (1984:

[30] See Mele (2005) for further reasons to reject negative events.

[31] Beebee (2004) offers another solution. Beebee proposes negative causal explanations provide information about the causal structure of the closest possible worlds where the events that failed to occur in the actual world did occur. So "Don did not die because the rope did not break" would tell us about the causal sequence that would have resulted had Don's rope broken. Negative causal explanations thus provide modal information. This solution is compatible with the Davidsonian view of causal explanations, although, as Beebee admits, it also does not prove that the Davidsonian view must be correct.

299). However, it seems odd to me to suggest that the truth of a negative causal explanation should depend on causal relations between events that take place somewhere else (perhaps even on causal relations between events that take place in non-actual possible worlds, because, even if no rope-breakings had ever occurred, and so no-one had ever died as a result of one, "Don did not die because his rope did not break" could still be true, and a Davidsonian might say this is because if some rope-breakings had occurred, these events *would* have caused deaths). It seems to me that the truth of negative causal explanations should depend on something within the causal system the causal explanation concerns. So, for example, "Don did not die because his rope did not break" should depend, for its truth, on Don, or something about Don—or the rope, or something about the rope. This is not a decisive objection against the response Child gives on behalf of the Davidsonian. Indeed, of the four kinds of counterexample I discuss in this chapter, negative causal explanations seem to me to be the least problematic for the Davidsonian view. However, it does highlight a cost of the Davidsonian view: on the Davidsonian view some causal explanations are made true by causally related events that occur outside the circumstances the causal explanation specifically concerns.

7.1.2 Process-citing explanations

A second group of counterexamples to the Davidsonian view is causal explanations that cite the continuous operation of causal processes, such as:

(d) The snow is melting because the sun is shining.

Are causal explanations like (d) made true by causally related pairs of events? As Alexander Mourelatos (1978) argues, process predications, of which "the snow is melting" and "the sun is shining" are examples do not implicitly quantify over events. So, (d) does not say that a melting event was caused by a shining event. The tense of (d) indicates that melting and shining are still going on, so it is not completed events but ongoing processes that the explanation references. Nevertheless, it may well be true that whenever the sun melts some snow by shining on it causal relations between events always obtain. For example, it might be that whenever the sun melts some snow by shining on it a series of causally related chemical events involving light particles and ice molecules occur. Perhaps it is *these* causally connected events on which the truth of (d) depends.

In most cases, when we say some causal process is in operation, we can find pairs of causally related events occurring at a finer temporal resolution. However, the vocabulary that we use to express the original causal explanation does not indicate what pairs of causally related events we should expect to find. For example, it is not part of the *meaning* of 'shining' or 'melting' that instances of

shining or melting involve causally related pairs of events of certain types.[32] It might be necessary that whenever the sun melts some snow by shining on it a series of causally related chemical events involving light particles and ice molecules occur, but this is an *a posteriori* necessity. The idea that an explanation must be made true by causally related events falling under types which have no connection to the *meaning* of the predications featuring in the explanation seems contrary to the reasonable principle that whatever makes some sentence true should be what the sentence is about. The notion of what a sentence is about is imprecise. Possibly, a Davidsonian could argue that, on a loose enough definition of 'aboutness', (d) is about events involving light particles and ice molecules. However, for this response to work, the Davidsonian would have to convince us to adopt his loose definition of 'aboutness'.

If one thought, as seems reasonable, that explanations are causal if and only if they answer to causal reality, and that all there is to causal reality is events standing in causal relations to other events, then it would be natural to suppose that (d) *must* depend for its truth on causally related pairs of events, if it is a causal explanation at all. However, as I argued in Chapter 6, one need not think of causation as always, everywhere a relation between events. Causation can be a determinable process engaged in by substances. If this view of causation is plausible, then facts about what events are causally related to what others are not the only causal facts that causal explanations could answer to. Some causal explanations may answer to facts about dynamic states of affairs. Furthermore, the idea that (d) is made true by facts about a dynamic state of affairs has intuitive appeal. What seems to matter for the truth of (d) is that it is the sun that is causing what the snow is suffering.

7.1.3 Stative causal explanations

A third group of counterexamples to the Davidsonian view are stative causal explanations. Here are three examples:

(e) The bridge collapsed because the bolt was weak. (Child 1994: 106)
(f) The floor is dirty because Mary's dog was here.
(g) My leg is broken because I fell off my bike. (Child 1994: 105)

These examples are problematic for the Davidsonian view because in each of them either the explanans clause or the explanandum clause, or both, seems to reference a state, not an event. In (e), that an event occurred is explained by the fact that a state obtains; in (f), that one state obtains is explained by the fact that another state obtained; and, in (g), that a state obtains is explained by the fact that an event occurred.

[32] Child (1994: 108) makes a similar point.

Once again, the Davidsonian can respond by stressing that reference to the events, whose causal connectedness grounds the truth of the causal explanation, can be concealed. The reply would go like this: when we talk of a state as the cause of some event, 'there is a causal relation between events; the state [is] part of the circumstances in which the cause occurred; and mentioning that state can help to explain why the cause had the effect it did' (Child 1994: 106). So, in the case of (e), something happened to cause the collapse of the bridge (e.g. a train went over the bridge); the bolt's being weak was part of the circumstances in which this event occurred and helps explain why the event caused the collapse of the bridge. Similarly, when someone offers "the floor is dirty because Mary's dog was here" as a causal explanation, we can suppose that events occurred that stand in causal relations to each other (e.g. Mary's dog arrived, then ran around the room with muddy feet, and this latter event caused the floor to become dirty) and these causally related events are what makes the stative causal explanation true. And, in (g), the causal explanation is made true by the causal relation obtaining between my falling off my bike and my leg breaking.

However, to suppose that whenever we offer a stative causal explanation there *must* be appropriate pairs of causally related events to serve as the grounds for the stative causal explanation seems to me to be metaphysically suspect. Events are not included in our ontology for the sole reason that they serve as truth-makers for causal explanations. Whether or not certain events exist and stand in causal relations, and whether or not a certain stative causal explanation is true, can therefore be determined independently. 'Was there an event that triggered the collapse of the bridge?' and 'did the bridge collapse because the bolt was weak?' seem like independent questions, in the sense that an answer to the first need not impact an answer to the second and vice versa. Confidence in the truth of the stative causal explanation should not, therefore, govern the truth of a claim about what events exist. Steward (1997: 173–174) also questions the assumption that appropriate pairs of causally related events can always be found to serve as the grounds for a stative causal explanation. In the bridge case, for example, what if the bridge just collapsed, apparently spontaneously? Must we always assume there was a triggering event that stands to the event explained as cause to effect?

7.1.4 Disposition-citing explanations

Stative causal explanations for which Steward's point seems particularly pertinent are stative explanations that seem to cite powers or dispositions. Indeed, (e) probably counts as a disposition-citing explanation. Other examples of disposition-citing explanations include:

(h) Peter sneezed because he is allergic to flowers.
(i) The cat died after eating the lilies because they are poisonous to cats.
(j) The aspirin relieved Joe's pain because it is a cyclo-oxygenase inhibitor.

It is possible that all stative causal explanations are disposition-citing explanations. For example, if it could be argued that (1) all stative predications attribute properties, and (2) all properties are really powers or dispositions, then it would follow that all stative causal explanations are really disposition-citing causal explanations. However, both of these premises are controversial.[33] I will not attempt to establish that all stative causal explanations are really disposition-citing explanations but I will assume that *some* stative causal explanations are disposition-citing explanations. I will also assume that disposition-citing explanations are causal explanations. As John Hyman puts it:

> [E]xplanations that refer to disposition are *echt* causal explanations, whatever kind of disposition they refer to. *How* they explain, exactly what part of a causal story they tell, and whether a disposition is the cause, or part of the cause, of its manifestation—these are contentious questions. But *that* explanations that refer to dispositions are causal explanations should be beyond doubt. (2015: 121)

Do disposition-citing explanations depend for their truth on the obtaining of causal relations between events? One might think that disposition-citing explanations are causal because they report causal relations between the triggering or stimulus event of the manifestation and the manifestation event. So, for example, perhaps (h) "Peter sneezed because he is allergic to flowers" reports a causal relation between Peter moving near to a flower (the trigger event) and Peter's sneeze (the manifestation event). For many dispositions, when they are manifested, causal relations between trigger and manifestation exist. Indeed, if they did not we might wonder whether the disposition has really been manifested at all. If there were no causal relation between Peter's moving near a flower and his sneeze, we might doubt that his sneezing was really a manifestation of his allergy. This is because to have an allergy is to be liable to exhibiting an immune reaction in the presence of an allergen—it is part of the meaning of 'allergy' that allergic reactions have specific triggers.

However, there are two problems with this suggestion. First, some dispositions do not seem to have triggers at all, either because they are always manifested (e.g. the disposition of a massive body to deform space-time) or because their manifestation is spontaneous (e.g. radioactive decay). Explanations that make reference to these sorts of dispositions therefore will not be made true by causal relations between triggers and manifestations, and, on the assumption that all disposition-citing explanations have the same sort of truth-maker, this casts doubt on the idea that disposition-citing explanations are made true by trigger-manifestation causal relations. Second, it is possible for there to be a causal relation between two events, the first of which is of

[33] Mumford (2004), Shoemaker (1980) and Whittle (2008) are three philosophers who have defended (2); Armstrong (1997: 69–84) has argued against it.

the same type as the trigger of a disposition's manifestation and the second of which is of the same type as a disposition's manifestation, without the disposition being manifested at all. For example, suppose the flower Peter moves near is bright white in colour, and the bright light reflected off the flower induces a photic sneeze reflex in Peter and he sneezes. In this example, moving near the flower caused Peter to sneeze, but his disposition to exhibit an immune response to flowers wasn't manifested. For all dispositions where the manifestation of a disposition involves a series of causally related events starting with a triggering event and ending with a manifestation event, it is possible for this type of causal chain to obtain without the disposition being manifested, because the causal chain is 'deviant' in some way.[34] This throws into doubt the idea that causal relations between trigger events and manifestation events are what disposition-citing explanations report.

One might think that disposition-citing explanations are made true by causal relations holding between the dispositions themselves and the events explained. However, I reject this suggestion because I do not think that dispositions or powers can be causal relata. A number of philosophers have doubted that dispositions or powers themselves can be causally efficacious. Debate about the causal efficacy or causal relevance of dispositions mirrors the debate about the causal efficacy or causal relevance of mental states. Frank Jackson (1995: 257) argues that, because part of what it is for a substance to possess a disposition, like 'fragility', is for that substance to be prone to exhibit the manifestation behaviour, this entails that the disposition is non-contingently connected to the manifestation behaviour. And, because the connection between cause and effect is contingent, this entails that the connection between disposition and manifestation cannot be causal. This parallels Abraham Melden's (1961: 52) objection to the idea that desires are causes of actions: desires are non-contingently related to actions that satisfy the desire. Elizabeth Prior, Robert Pargetter and Frank Jackson (1982) argued that dispositions lack causal efficacy because there is always a 'causal basis' of the disposition—i.e. there is always a 'property or property-complex of the object that, together with the [triggering or stimulus event] is the causally operative sufficient condition for the manifestation in the case of "surefire" dispositions, and in the case of probabilistic dispositions is causally sufficient for relevant chance of the manifestation' (1982: 251). According to Prior, Pargetter and Jackson, this means that there is no 'causal work' left for the disposition to do (unless the manifestation event is overdetermined). This argument parallels Jaegwon Kim's causal exclusion argument, discussed in Chapter 1. And, just as philosophers have responded to Kim by questioning assumptions about what it means for a mental property or state to be causally relevant, philosophers have responded to Prior, Pargetter and Jackson by questioning assumptions about what it means for a disposition be causally relevant (e.g. McKitrick 2005).

[34] Hyman argues for this point (2015: 121–127).

However, I think that the debate about the causal efficacy or causal relevance of powers/dispositions is often misconceived. In Chapter 6, I expressed support for the Rylean view that powers are not *things*; they are not 'elements of being', to borrow a phrase from E. J. Lowe (2005). In Ryle's view, to attribute a power to an entity is not to report a state of affairs; it is not to say that the entity has some attribute or stands in some relation. For an entity to have a power is for an open-ended set of facts about what that substance can do, or can be relied upon to do—what processes it can engage in—to be true of it. Powers are ways of thinking about how substances are connected to the processes they engage in. In this respect, *power* is a concept that does not name any kind of being but instead helps us explain the ontological form of entities belonging to the categories the concept concerns. If this view is correct, and for a substance to have a power is not for it to have a certain attribute or stand in a certain relation, then powers (or the state of having a power) cannot be relata of *any* relation, let alone a causal relation. Arguments like Prior, Pargetter and Jackson's only have bite if one assumes that powers are the sorts of entities that even could 'do causal work'—and I do not think powers or dispositions are the sorts of entities that even could 'do causal work', because I do not think they are any sort of entity at all.

If one thought that causal reality were nothing but events standing in causal relations, then explanations that make reference to dispositions, if they are causal at all, would have to depend for their truth on the obtaining of certain types of causal relations. However, if the non-relationalist view of causation put forward in Chapter 6 is plausible, then causal reality is more than events standing in causal relations to other events; it is also a matter of substances engaging in processes. The idea that it is something about this latter aspect of causal reality that disposition-citing explanations answer to is plausible. On the non-relational theory of causation I outlined in Chapter 6, what it is for a substance to be exercising a power, or manifesting a disposition, is for that substance to be engaging in a process. Therefore, the obvious candidate for what a disposition-citing explanation reports is the fact that some dynamic state of affairs is a manifestation of the disposition cited. In other words, disposition-citing explanations depend for their truth on the relationship between the disposition cited and the dynamic state of affairs that is the manifestation of that disposition.

7.2 Causal explanations and manipulation

We have seen that some causal explanations—namely negative causal explanations, causal explanations that cite the operation of causal processes, stative causal explanations, and disposition-citing causal explanations—do not explicitly mention events whose causal connectedness could ground their truth. In the face of causal explanations like this, the Davidsonian is forced to maintain that reference to the causally related events that make true a causal explanation

can be opaque. This suggestion is not implausible itself, but in the case of negative causal explanations and causal explanations that cite the operation of causal processes it threatens to contravene the reasonable assumption that what makes a sentence true must be what the sentence is about. Furthermore, even this response seems insufficient in the case of stative causal explanations and disposition-citing explanations. This is because, for at least some stative causal explanations and disposition-citing explanations, it is not obvious that causally related pairs of events can be found to serve as implicit referents of explanandum and explanans.

Child suggests that, in the face of counterexamples like those discussed in Section 7.1, we could 'give up the idea that what makes an explanation a causal explanation is its dependence on the presence of causal relations between events' (1994: 109). There is more than one way to do this. First, we can give up this idea *without* giving up the idea that what makes an explanation causal is its dependence on the presence of causal relations of some other kind (perhaps between states). Second, we can deny that what makes an explanation a causal explanation is its dependence on the presence of causal relations of *any* kind—what unites causal explanations into a single category is something else, perhaps a fact about the sort of information they provide.

Some of Child's remarks suggest that he has sympathy for the second option. He describes the alternative to the Davidsonian account as a view where 'causal explanations are not united by their dependence on a natural relation of causality, but rather by the fact that they are all explanations of the occurrence or persistence of particular events or circumstances, or of general types of event or circumstance' (1994: 100). In any case, it should be obvious that I prefer the second option. I concede that causal explanations depend for their truth on an underlying causal reality, but this underlying reality need not involve any causal *relations*—some causal explanations are not grounded by the presence of any causal relation at all. Instead, I think that explanations are causal because of the sort of information they provide.

In Chapter 6, I discussed an objection to my view that we think of causation in two distinct ways, as a process and separately as a cause–effect relation. According to this objection, the idea that we think of causation in two different ways is inconsistent with the idea that causation is a single phenomenon. I responded to this objection by maintaining that the concept of causation as a cause–effect relation is derived from our concept of causation as a process that substances engage in. I noted that, if substances possess and exercise causal powers, then substances with certain powers must behave in certain ways when the conditions for the manifestation of the power arise, provided there is nothing interfering. The exercise of powers will therefore be the source of regular and stable relations between trigger events and manifestation events. We can use knowledge of these relations to change how powerful substances behave. For example, if one knows that being near flowers triggers an allergic reaction, then one can prevent the allergic reaction by avoiding flowers.

From this we get the idea that events, particularly (but not exclusively) trigger events, can be *devices* for manipulating later events. Events are not literally devices but, even though talk of events as devices is metaphorical, there are still conditions under which use of this metaphor is correct and conditions under which use of this metaphor is incorrect. The metaphor is thus the source of the idea that there is a special sort of relation between events, which is causation. So, the causation concept can cover ontologically diverse phenomena, because from the concept of causing as something substances engage in, we can derive the idea that some relations between events are causal, via the intermediary notion of using knowledge of stable relations between trigger events and manifestation events to manipulate powerful substances.

The notion of manipulation thus ties the concepts of causation as a process and causation as a relation together. I suggest that the notion of manipulation is also what explains how many diverse explanations can all count as causal. Causal explanations are those that provide information relevant to the manipulation of an effect. They are explanations that provide us with information about how to stop something from happening, or how to get something to happen again, or how to get it to happen in a different way (or at least information about how to make such outcomes more likely). These criteria for an explanation to be causal are similar to criteria suggested by Bradford Skow (2013). Skow claims that 'A body of facts partially causally explains E if it is a body of facts about what causes, if any, E had; or if it is a body of facts about what it would have taken for some specific alternative or range of alternatives to E to have occurred instead' (2013: 449). Skow defends this theory of causal explanation on the grounds that there are many explanations that provide causal information but which do not name an event that stands in a causal relation to the explanandum.

One might argue that my proposal gives conditions that are unnecessary for an explanation to be causal, because there are some causal explanations where the named causal factor cannot be manipulated even in principle. For example, one might think that "Fido is warm-blooded because he's a dog" and "Sarah didn't get promoted because she's a woman" are causal explanations.[35] It is impossible to consider whether or not Fido would have been cold-blooded had he not been a dog, because any possible being that is not a dog is not Fido; similarly, it is impossible to consider whether or not Sarah would have got promoted had she not been a woman, because any possible being who is not a woman is not Sarah, or so the thought goes. For this reason, these cannot be examples of explanations that give information relevant to the manipulation or control of an effect.

In response to the first example, it is not obvious to me that this explanation is a causal explanation at all. Fido's being warm-blooded is not *causally* explained by his being a dog—being warm-blooded is part and parcel of what it is to be

[35] Holland considers examples of this kind, arguing that if these really are causal claims then they are causal claims that lack a clear meaning (1986: 954–956).

a dog. The second example, in contrast, does seem to me to be a causal explanation. However, it is not obvious that Sarah's gender is an essential property of her, so it is not obvious that any possible being who is not a woman is not Sarah. Furthermore, even if Sarah's gender were an essential property of her, I would argue that social categories like gender, race and class (and perhaps also categories like criminal, employee, preacher, grandmother etc.) are peculiar in that the dispositional properties one enjoys or suffers as a result of being placed into one or other of these categories only exist because of certain cultural practices and behaviour. Sarah's being a woman is a causal factor in the explanation of her not getting promoted, but only because, as a society, we are liable to treat people differently when they fall into different social categories. So, even granting that Sarah's gender is not, even in principle, something we can manipulate, the cultural practices and behaviours that turn being a woman into a causal factor in the first place are certainly things we can manipulate. In other words, "Sarah didn't get promoted because she's a woman" is an explanation that provides information relevant to manipulation of an effect after all, because of the peculiar connection between social categories and changeable cultural practices. Of course, exactly how social categories function is a debated topic, but this only emphasises the point that "Sarah didn't get promoted because she's a woman" is not an uncontroversial counterexample to my proposal.[36]

In this chapter, I have sought to show that it is not obviously true that an explanation is causal only if its explanandum designates an effect and its explanans designates an item that is the cause of that effect. My non-relational theory of causation allows that some causal explanations may depend for their truth on facts about dynamic states of affairs. Furthermore, it is quite plausible that process-citing explanations and disposition-citing explanations are the kinds of causal explanation that answer to the non-relational aspect of causal reality. In other words, it is plausible that the 'because' of these causal explanations does not signify the obtaining of a causal relation.

References[37]

Armstrong, D M 1997 *A world of states of affairs*. Cambridge: Cambridge University Press.

Beebee, H 2004 Causing and nothingness. In: Collins, J, Hall E J and Paul, L A, *Causation and counterfactuals*. Cambridge, MA: MIT Press. pp. 291–308.

Child, W 1994 *Causality, interpretation, and the mind*. New York: Oxford University Press.

[36] See Woodward (2003: 114–117) for a good discussion of this issue.

[37] Author note: some references to Davidson are formatted (1963/2001). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001 collection of his essays, with the page numbers relating to that volume.

Clarke, R 2014 *Omissions: Agency, metaphysics, and responsibility*. New York: Oxford University Press.

Davidson, D 1963 Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700. DOI: https://doi.org/10.2307/2023177. Reprinted in Davidson 2001 pp. 3–20.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Holland, P W 1986 Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960. DOI: https://doi.org/10.2307/2289064

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Jackson, F 1995 Essentialism, mental properties and causation. *Proceedings of the Aristotelian Society*, 95: 253–268. DOI: https://doi.org/10.1093/aristotelian/95.1.253

Lowe, E J 2005 *The four-category ontology: A metaphysical foundation for natural science*. Oxford: Clarendon Press.

McKitrick, J 2005 Are dispositions causally relevant? *Synthese*, 144(3): 357–371. DOI: https://doi.org/10.1007/s11229-005-5868-z

Melden, A I 1961 *Free action: Studies in philosophical psychology*. London: Routledge & Kegan Paul.

Mele, A R 2005 Action. In: Jackson, F and Smith, M *The Oxford handbook of contemporary philosophy*. Oxford: Oxford University Press. pp. 78–88.

Mourelatos, A 1978 Events, processes and states. *Linguistics and Philosophy*, 2(3): 415–434. DOI: https://doi.org/10.1007/bf00149015

Mulligan, K, Simons, P and Smith, B 1984 Truth-makers. *Philosophy and Phenomenological Research*, 44(3): 287–321. DOI: https://doi.org/10.2307/2107686

Mumford, S 2004 *Laws in nature*. London: Routledge.

Prior, E, Pargetter, R and Jackson, F 1982 Three theses about dispositions. *American Philosophical Quarterly*, 19(3): 251–257.

Shoemaker, S 1980 Causality and properties. In: van Inwagen, P *Time and cause*. Dordrecht: D. Reidel. pp. 109–135.

Skow, B 2013 Are there non-causal explanations (of particular events)? *British Journal for the Philosophy of Science*, 65(3): 445–457. DOI: https://doi.org/10.1093/bjps/axs047

Steward, H 1997 *The ontology of mind: Events, processes and states*. Oxford: Oxford University Press.

Strawson, P F 1985 Causation and explanation. In Bruce Vermazen, B and Hintikka, M B *Essays on Davidson: Actions and Events*. Oxford: Oxford University Press. pp. 115–135.

Whittle, A 2008 A functionalist theory of properties. *Philosophy and Phenomenological Research*, 77(1): 59–82. DOI: https://doi.org/10.1111/j.1933-1592.2008.00176.x

Woodward, J 2003 *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

CHAPTER 8

Action Explanation

In the previous chapter, I argued that it is not necessary for an explanation to be causal that its explanandum designate an effect and its explanans designate an item that is the cause of that effect. My non-relational theory of causation implies that facts about causal relations between particulars are not the only causal facts that could ground the truth of causal explanations. I suggested that some causal explanations are made true by the non-relational aspect of causal reality, that is, by facts about substances engaging in processes. In this chapter, I turn my attention back to rationalising explanations of action. Rationalising explanations of action explain why an agent acted as she did (this is the explanandum) by telling us why, in the agent's eyes, what they did was a rational thing for them to do (this is the explanans). I will argue that rationalising explanations are also causal explanations that are not made true by a pair of causally related events.

The debate concerning how we ought to understand rationalising explanations is central within philosophy of action because part of what makes intentional actions distinctive is that often when we explain an intentional action we do so by giving the agent's reason for acting. The nature of intentional action is thus inseparable from their appropriateness for receiving rationalising explanations. Whatever intentional actions are, they must be things that can be explained by reasons.

For a long time, opinion on rationalising explanations has been divided in two. There are those who endorse the causal theory of action explanation, which says that rationalising explanations explain by giving a *causal* account of the agent's action, and moreover that 'the primary reason for an action is its cause' (Davidson 1963/2001a: 4). Then there are non-causalists, who believe that the concepts cited in rationalising explanations, like *belief, desire* and *intention*, do not refer to items that can stand in causal relations to actions, so, when

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation.* Pp. 155–174. York: White Rose University Press. DOI: https://doi.org/10.22599/White.i.

such concepts are employed to explain why an agent acted, they do not explain by giving a *causal* account of the agent's action.

The causal theory of action explanation is widely supported in large part due to what has become known as 'Davidson's challenge', which runs as follows. Some statements that tell us why what an agent did seemed to them to be rational do not explain why the agent did as she did. This kind of statement is a 'mere rationalisation'. Mere rationalisations are similar to rationalising explanations in that they tell us why the course of action taken by the agent seemed, to the agent, to be a rational course of action to take. However, mere rationalisations do not tell us why an agent acted as she did. If Anna ends up speaking at the conference to impress her friends and not because it will be good for her career, even though she considers being good for her career to be a sound reason to speak at the conference, then (a) is a mere rationalisation of her action.

(a) Speaking at the conference seemed rational to Anna because it would be good for her career.

This is a mere rationalisation because it explains why speaking at the conference seemed to Anna to be a rational thing for her to do—but it does not explain why Anna actually spoke at the conference. It is not true that Anna spoke at the conference *because* she thought it would help her career. That she wanted to impress her friends, on the other hand, *does* explain why Anna acted as she did. (a') is a genuine rationalising explanation of Anna's action.

(a') Anna spoke at the conference because she wanted to impress her friends.

Because some rationalisations do not explain why the agent did as she did, those rationalisations that do—like (a')—must achieve this by doing more than simply revealing why what an agent did seemed to them to be a rational thing to do. And if the extra thing that rationalising explanations do is not revealing causal information, then what is it?

As mentioned in Chapter 2, Jonathan Dancy argues that the difference between rationalising explanations and mere rationalisations is just that the former tell us 'the considerations in the light of which he acted' and the latter only tell us 'considerations he took to favour acting as he did but which were not in fact ones in the light of which he decided to do what he did' (2000: 163). On Dancy's view, 'acted in the light of' performs the function in the case of rationalising explanations that truth plays in the case of other sorts of explanation: it makes the difference between a statement that explains and a statement that does not explain. Dancy thinks the fact that 'acted in the light of' can perform this function is a brute fact. However, while I think it is plausibly a brute fact that only true statements can explain, there is something perplexing about the fact that 'acted in the light of' can perform the same sort of function truth can. Why does 'acted in the light of' bestow explanatory power? What pattern of

explanation is demonstrated when we give the reason an agent acted in the light of? For Davidson, the answer is the sort of explanatory pattern we see when we give the cause of an effect. According to Davidson's proposal, of all the reasons Anna had for speaking at the conference, that it would impress her friends is the reason that explains her action because this reason somehow identifies the cause of her action. As emphasised in Chapter 2, Davidson's view is not just that rationalising explanations explain by giving causal information—they also identify *the cause* of the action explained.

It is important to note that, even though Davidson claimed that 'the primary reason for an action is its cause' (1963/2001a: 4), strictly speaking Davidson did not think beliefs or desires were causes. Believing that something is the case and desiring to do something are not events but states. In (a') the explanans, i.e. 'she wanted to impress her friends', would be classified by Mourelatos as a state predication, not an event predication. On the assumption that causal explanations are typically explanations that tell us what event stands in a causal relation to the event whose occurrence we want to explain, the fact that the explanans of most rationalising explanations is a state predication seems to speak against classifying these explanations as causal. This difference between rationalising explanations and typical causal explanations, however, does not carry much force. Davidson acknowledged that beliefs and desires are not events but states—in fact, he thought that they were 'dispositions to behave in certain ways' (1997/2001b: 72)—and as such could not literally be causes. However, Davidson argued that *the onset of* a belief and *the onset of* a desire are events, and explanations of actions that cite beliefs and desires are explanatory if the belief or desire attribution is 'closely associated' (1963/2001a: 12) with an inner mental event such as the onset of the belief or desire that is the cause of the action explained. As Davidson puts it, 'In many cases it is not difficult at all to find events very closely associated with the primary reason. States and dispositions are not events, but the onslaught of a state or disposition is' (1963/2001a: 12). The causal relation that makes the rationalising explanation explanatory does not need to be *explicitly* reported by the rationalising explanation. As explained in the previous chapter, the Davidsonian position on stative causal explanations is that, even though states cannot literally be causes, states can be 'part of the circumstances in which the cause occurred; and mentioning that state can help to explain why the cause had the effect it did' (Child 1994: 106). Davidson allowed the relation between a causal explanation and the causally related events that make the explanation true to be opaque (Child 1994: 106), which means that reference to the events whose causal connectedness grounds the truth of a rationalising explanation need not be transparent.

Another nuance of Davidson's theory of rationalising explanations concerns how he deals with the following issue. When we causally attribute one event to another, this is usually taken to imply the existence of a law that states that there is an event-kind F, of which the cause event is a token, and an event-kind G, of which the effect event is a token, such that F events always cause G events.

However, when we say that an agent acted as she did because of the beliefs and desires she had, there is no implication that other agents with the same beliefs and desires will (or are likely to) do the same thing, or that the same agent will act in the same way when she has the same beliefs and desires on another occasion (Hart & Honoré 1985: 55). Davidson's anomalous monism allows him to concede this point without giving up the idea that rationalising explanations must be causal explanations. Davidson proposes that when a mental event and an action are causally related, these two events fall under event-kinds that feature in a causal law. This follows from Davidson's principle of the nomological character of causality: all causal relations are covered by strict deterministic laws. However, the event-kinds that feature in the causal law, which the mental event and action fall under, are *physical* kinds, not mental kinds. Furthermore, the law that covers the causal relation can *only* be stated in a language of physical kinds. Rationalising explanations do not imply lawlike regularities between mental states and actions because in giving a rationalising explanation we are picking out the cause of an action using mental kinds, and these mental kinds do not feature in any universal regularity, not even the universal regularity that covers the causal relation that the rationalising explanation owes its success to. As Davidson puts it:

> The laws whose existence is required if reasons are causes of actions do not, we may be sure, deal in the concepts in which rationalisations must deal. If the causes of a class of events (actions) fall in a certain class (reasons) and there is a law to back each singular causal statement, it does not follow that there is any law connecting events classified as reasons with events classified as actions—the classifications may even be neurological, chemical, or physical. (1963/2001a: 17)

Consequently, if anomalous monism is true, we should expect that when we say that an agent acted as she did because of the beliefs and desires she had, there is no implication that other agents with the same beliefs and desires will (or are likely to) do the same thing, but this is because in giving a rationalising explanation we are picking out the cause of an action using *mental* kinds, and these mental kinds do not feature in any universal regularity.

There has recently been renewed interest in developing a non-causal account of rationalising explanations that meets Davidson's challenge. For example, Scott Sehon (2005) argues that rationalising explanations are *teleological explanations* that are irreducible to causal explanations; Julia Tanney (2009; 2013) argues that rationalising explanations are *context-placing explanations*; and Megan Fritts (2021) argues that action explanations are *structural explanations*.

In order to determine whether any of these non-causal accounts meet Davidson's challenge, it is useful to outline what the success criteria for meeting Davidson's challenge are. It is important to recognise that Davidson's challenge is not an epistemological challenge: it is not a question of how we know

which reason an agent acted in the light of. For example, Davidson is not challenging us to explain *how we know* that Anna spoke at the conference because she wanted to impress her friends and not because it would be good for her career. For Davidson, this epistemological question is answered by considering how the reason fits with the agent's general character, how the reason coheres with the agent's other beliefs and desires, whether the reason reveals the agent to be acting rationally etc. However, Davidson stresses that 'it is an error to think that, because placing the action in a larger pattern explains it, therefore we now understand the sort of explanation involved' (1963/2001a: 10). Davidson's challenge is a theoretical challenge. It can only be successfully answered with a satisfying theory of how rationalising explanations explain. Successfully meeting Davidson's challenge requires answering the following question: what is it about the reason we know is the reason for which the agent acted that *qualifies* the connection between reason and action as an explanatory connection?

The view I advance in this chapter falls somewhere in between the Davidsonian and the non-causalist view. Like most non-causalists, I agree that the concepts cited in rationalising explanations do not seem to discharge their explanatory function by denoting (even opaquely) causes of the actions they explain. However, I do not think non-causal accounts of rationalising explanations successfully meet Davidson's challenge—at least, the non-causal theories that have the best chance of meeting Davidson's challenge are also theories whose classification as 'non-causal' is questionable. The way forward is, I think, to adopt a position that sits somewhere in between non-causalism and causalism. First, in Section 8.1, I will examine considerations that motivate seeking a non-causal account of rationalising explanations. In Section 8.2 I will assess some non-causalist accounts of rationalising explanations. I will argue that the 'non-causalist' accounts of rationalising explanations that stand the best chance of meeting Davidson's challenge could be considered 'causal' after all if one takes a non-Davidsonian approach to what makes an explanation causal. In Section 8.3 I will make the case for thinking that rationalising explanations are causal explanations, but causal explanations that are unique in two important ways.

8.1 Mental concepts

Rationalising explanations display certain features that set them apart from typical causal explanations, like 'the patient developed cancer because he was exposed to radiation'. As already mentioned, many rationalising explanations explain an action by reference to a state of the agent, as opposed to an event involving the agent. For example in (a')–(c) the explanantia are state predications:

(a') Anna spoke at the conference because she wanted to impress her friends.
(b) Beth is buying flour because she wants to make bread.

(c) Carlin is adding rosemary to the sauce because he believes it will make it taste better.

Many rationalising explanations offer facts as explanans:

(d) Daniel took the A road because the motorway was shut.

Some rationalising explanations include infinitival phrases as explanations:

(e) Esther is breaking eggs to make an omelette.

Many rationalising explanations, which Michael Thompson (2008) calls 'naïve action explanations', explain one action in terms of another:

(f) Fred is drilling a hole in the wall because he is hanging a picture.

Some rationalising explanations explicitly mention intentions or what the agent is trying to do. However, because rationalising explanations are so variable in form, not very much at all can be concluded about how rationalising explanations explain by considering the language of rationalising explanations. Furthermore, a core tenet of the Davidsonian view is that the events whose causal connectedness grounds the explanatory power of a rationalising explanation need not be *explicitly or transparently* referenced in the rationalising explanation. However, one motivation for seeking a non-causal theory of rationalising explanation is that the concepts employed in rationalising explanations, such as *belief, desire, intention, goal* and *attempt*, do not seem to discharge their explanatory role by designating a cause of the action they are invoked to explain—not even implicitly or opaquely. The point here is that rationalising explanations do not seem to have anything to do with finding the cause of an action.

Elizabeth Anscombe points out that, when 'one says what desire an act was meant to satisfy, one does not identify a feeling, image or idea that precedes the act the desire explains' (2000: 17). The desire that an act satisfies is not the 'mental cause' of the act in the same way that, to use Anscombe's example, noticing a face appearing at the window might be the mental cause of one's jumping. Anscombe defines a 'mental cause' as 'what someone would describe if he were asked the specific question: what produced this action … on your part: what did you see or hear or feel, or what ideas or images cropped up in your mind, and led up to it?' (2000: 17–18). Giving a 'mental cause' of something, in the special sense of 'mental cause' that Anscombe has isolated, is thus to say what prior mental event *triggered* one's action. Rationalising explanations do not seem to be like this: they do not seem to be explanations whose explanatoriness depends on them identifying, or at least suggesting, the event that triggered the action. When we explain actions by citing an agent's beliefs or desires, we are usually not identifying something that occurred at a particular

time that triggered the action, or which moved the agent from a state of inaction to a state of action. When we seek a rationalising explanation of someone's action, what triggered the action, what event made the difference to its occurring, does not seem to matter.

Julia Tanney makes a similar point. Tanney argues that the concepts that are at work in rationalising explanations perform their explanatory role 'without designating *anything*; let alone causally efficacious states or events; let alone causally efficacious states or events whose nature awaits discovery' (2009: 100, emphasis added). Tanney claims that assuming that mental concepts designate 'logically independent, temporally antecedent, causally efficacious events' is to assume that mental concepts are 'theoretical terms' (2009: 100). A theoretical term is one that purports to refer to an event, property, state, fact or condition whose intrinsic nature is up for discovery but which causes a phenomenon to be explained. An example of such a theoretical term would be 'gene': genes are entities we posit on the grounds that their existence would explain some observable phenomena; 'gene' is a term that purports to refer to a hidden but causally efficacious entity. Tanney argues that treating mental concepts as theoretical terms 'mis-assigns the explanatory function of these concepts' (2009: 100).

The position commits us to postulating an event, unobservable to others and possibly even to the agent herself, that would, if known, provide the sought-after reason-explanation for the agent's action. In such cases, as Ryle insists, an epistemological puzzle arises as to how anyone could ever know whether a person acts for reasons or what, if she does, her reasons are, since the hypothesis is not even in principle testable. Not only do we not, in everyday situations, have access to these hidden events, but even if we were, say, to monitor the neural activity of someone's brain or access their stream of consciousness, we would never be able to set up the kinds of correlations that would establish a particular occurrence as an instance of a particular reason without already having a way of deciding whether someone acted for a particular reason in order to make the correlation (Tanney 2009: 100).

Tanney's point here is that, if we construe mental concepts as designating hidden inner causes of behaviour, then rationalising explanations become the kind of claim whose truth depends on the existence of events we have no access to and that just does not seem to be how rationalising explanations work.

One might argue that Anscombe's and Tanney's view that when we explain each other's intentional behaviour we do not do so by positing inner mental causes that produce the behaviour is just an intuition. Jerry Fodor is one among many who has the opposite intuition (1987). However, Devin Curry (2018) summarises empirical evidence from experimental psychology that seems to support Anscombe's and Tanney's interpretation of what we are doing (or what we are not doing) when we give rationalising explanations.

Curry cites evidence from within the field of 'attribution theory', the branch of psychology concerned with how people explain each other's behaviour.

Research by Bertrum Malle indicates that there are important differences between the types of explanation people give for accidental behaviour and the types of explanation people give for intentional behaviour. When explaining accidental behaviour, people are more likely to reference 'inner causes', treating these as one would a mechanical cause of a physical event (Malle 2004: 61). The more intentional a behaviour appears to be, the more likely people are to explain the behaviour in terms of reasons (Malle 1999: 28–31). Curry also cites research that shows that children draw a distinction between *mistakes*, which demand explanation in terms of beliefs, and *accidents*, which demand explanation in terms of causes in the physical environment (Hatano & Inagaki 2002; Schult & Wellman 1997). Further research indicates that the types of questions people ask of someone's behaviour are different depending on whether the behaviour is perceived to be intentional or unintentional. Only for unintentional behaviour do people ask what *produced* the action (Malle, Knobe & Nelson 2007; Monroe & Malle 2017). From this research, Curry concludes that when it comes to explaining intentional behaviour people are concerned with placing the behaviour in a context that makes the behaviour understandable—people are concerned with identifying what produced behaviour only when the behaviour is perceived to be unintentional or accidental.

Interestingly, the psychological evidence is inconclusive with regard to whether people regard reason-citing explanations of intentional actions as causal or not. Curry writes that, even though people seem to treat reason-citing explanations of intentional behaviour as distinct from mechanistic explanations of accidental behaviour, previous work has shown that teleological explanations are often considered a kind of causal explanation (DiYanni & Kelemen 2005; Lombrozo & Carey 2006). What Curry's discussion shows is that Anscombe's intuition—that when we explain intentional behaviour we do not do so by identifying an 'inner cause' or 'mental trigger' that produced the action—cannot be easily dismissed. An assumed contrast between rationalising explanations and mechanistic explanations underlies much of our actual attributional behaviour. The empirical evidence Curry (2018) cites seems to suggest that we generally treat reason-citing explanations of intentional actions as distinct from causal or mechanical explanations, only using the latter kind of explanation when explaining unintentional or accidental actions.

Another observation that gives us reason to doubt that rationalising explanations explain by giving the cause of the explanandum is that sometimes when an agent has more than one reason for performing some action it is genuinely indeterminate which of the reasons was the reason she acted for. As Erasmus Mayr puts it, there is not always a fact of the matter about which reason an agent acted for. Consider cases where the agent has a bundle of strong motives to do X but it is not clear—even after thorough examination of his action, its circumstances and his general character—on which of these motives he has acted. We do not have to assume that our inability to decide this question rests

on merely practical grounds—that is, that there is a fact of the matter that we are unable to establish only because we lack further evidence—for it may well be that we would not even know what kind of further evidence would decide the question. Instead, we should accept that in such cases our inability may stem from the fact that these cases are truly indeterminate, because the criteria for judging whether the agent acted on a particular reason have 'run out' without unequivocally determining an answer (Mayr 2011: 261).

The idea that rationalising explanations explain by identifying the cause of the action is inconsistent with allowing for this kind of indeterminacy. On the Davidsonian view, any indeterminacy regarding what belief or desire the agent acted in the light of is epistemic—this is because an agent acts in the light of a belief or desire if and only if the onset of that belief or desire is the cause of the action, and the latter relation cannot be indeterminate. Of course, it could be that, when an agent has many reasons favouring a course of action, their action is causally overdetermined by these many reasons. However, it seems possible that an agent could have many reasons favouring a course of action, where none of these reasons is the reason the agent acted, and where the agent would not have acted if the case for acting was not overwhelming. For example, imagine Anna is again deciding whether or not to speak at a conference, and because the conference is quite far away Anna vows only to speak at the conference if the case for doing so seems overwhelming, where overwhelming for her means that there are at least *n* strong reasons favouring the action (where *n* is more than one). Then suppose Anna discovers *n* reasons for speaking at the conference, and so goes on to speak at the conference, but none of Anna's reasons stands out as the reason for which Anna spoke at the conference. In this case, it does not seem like Anna acts in the light of just one of the many reasons favouring speaking at the conference, but it is also not plausible to describe this as a case of overdetermination by her *n* reasons, because it is not the case that Anna would have acted in the same way had any one of her *n* reasons been missing.

8.2 Rationalising explanations as non-causal explanations

Rationalising explanations do not seem to explain action by designating inner causes of behaviour. However, the power of Davidson's challenge is that, if Davidson's answer is the only satisfactory answer to the challenge, then, regardless of how rationalising explanations *seem* to function, their explanatoriness *must* be grounded by causal relations between events somehow identified by mental concepts and the actions explained. What seems to matter, then, is whether there is a successful non-causal answer to Davidson's challenge, which is, as stated above, to explicate what is it about the reason we know is the reason for which the agent acted that *qualifies* the connection between reason and action as an explanatory connection. In the next two sections I will consider

two non-causal accounts of rationalising explanation that I think have the best chance of meeting Davidson's challenge.[38]

8.2.2 Context-placing explanations

Tanney suggests rationalising explanations ought to be understood as 'context-placing' explanations. The explanans of a rationalising explanation explains the action by placing it in a context that makes it intelligible. According to Tanney, rationalising explanations are explanations that work by giving us more information about what is going on. Tanney provides the following example of a 'context-placing explanation':

(g) The teacher has written 'CAT' on the board because she is writing 'CATALYST' on the board.

According to Tanney, the explanans does not illuminate 'any mysterious connection between the occurrences of two contingently related events—the writing of "c", "a", and "t", on the one hand and the writing of "catalyst", on the other' (2009: 98). Instead, the explanans in (g) 'serves to re-characterise what happened so that it—as newly described—is no longer puzzling' (Tanney 2009: 98). Tanney argues that rationalising explanations are all, essentially, of this kind.

Tanney's theory of rationalising explanations as context-placing is similar to a suggestion made by Anscombe that rationalising explanations 'interpret' the action explained:

> To give a motive … is to say something like "See the action in this light". To explain one's own actions by an account indicating a motive is to put them in a certain light. (2000: 21)

[38] The two accounts I will consider are not the only non-causal accounts of rationalising explanation available. Scott Sehon (2005) argues that rationalising explanations are *teleological explanations*, which are irreducible to causal explanations. On this account, rationalising explanations explain by making clear the aim the agent's behaviour was directed towards. Anna's action is directed towards impressing her friends, not improving her career prospects. That's why the former, but not the latter, explains her action. What makes it the case that Anna's action is directed towards impressing her friends and not towards improving her career prospects? Sehon argues that facts about what an agent is aiming to achieve are not reducible to causal facts; instead they are their own *sui generis* kind of teleological fact (see also: Löhrer & Sehon 2016; Sehon 2007: 163–165; Sehon 2010: 125). Unlike Sehon's account, the two accounts I will consider do not entail any substantial metaphysical commitments.

Mayr (2011: 269) also endorses the idea that rationalising explanations 'explain actions by making them intelligible' and not by positing an event-causal link between the agent's action and an appropriate mental event. For Mayr, rationalising explanations explain by providing us with a way of framing the agent's actions—a way of seeing the agent's actions as manifesting a certain pattern.

Tanney's view also bears some similarity to a view advanced by Michael Thompson (2008). Thompson (2008) outlines a class of rationalising explanations that he calls 'naïve action explanations'. These rationalising explanations explain one action in terms of another. (f) would be an example of such a rationalising explanation:

(f) Fred is drilling a hole in the wall because he is hanging a picture.

Tanney's paradigm context-placing explanation also explains why an agent engaged in some activity in terms of something else the agent is doing. Thompson suggests that most of the time when we explain our intentional actions we do so by citing another activity we are engaging in, of which the action to be explained is a part. Thompson grants that not all rationalising explanations have this form but he argues all rationalising explanations depend for their success on being suitably related to a relevant naïve action explanation. Tanney, Mayr and Thompson all seem to have hit upon what is essentially the same idea: that rationalising explanations explain by situating an agent's action within a wider pattern of activity the agent is engaging in which thereby makes the action expected.

How does this theory of rationalising explanations meet Davidson's challenge? Tanney says in cases where there are multiple reasons that could make sense of an agent's action but only one that genuinely explains why the agent acted as she did, we may simply need to 'probe further for a different or more far-reaching context-placing explanation that will succeed or give up the initial expectation that the action can be explained by reasons' (2009: 100). In other words, the distinction between a mere rationalisation and a genuinely explanatory rationalising explanation is that the latter, but not the former, fits better with a more far-reaching account of the agent's activities. So Anna spoke at the conference because she is trying to impress her friends, and not because it would be good for her career, because she is not trying to improve her career prospects. Her current action (speaking at the conference) is a constituent of her broader action of trying to impress her friends; it is not a constituent of trying to improve her career prospects because that is not something Anna is doing. This can be seen if we take a broad enough appraisal of Anna's activities and plans. Mayr would add that Anna spoke at the conference because she is trying to impress her friends and not because it would be good for her career because Anna is following standards of success set by impressing her friends and not standards of success set by improving her career prospects—this can be seen once we take into account the fact that Anna would be pleased if her

friends were impressed by her conference talk, disappointed if they were unimpressed, and feel nothing if her career prospects improved.

This issue with Tanney's (and Mayr's) response to Davidson's challenge is that the considerations outlined above seem only to address the epistemological question of how we know which reason an agent acted in the light of. Davidson's warning that 'it is an error to think that, because placing the action in a larger pattern explains it, therefore we now understand the sort of explanation involved' (1963/2001a: 10) seems particularly applicable to the context-placing view. What we need to know is why situating an agent's action within a wider pattern of activity the agent is engaging in qualities as an *explanation* of that action. What qualifies the connection between the agent's action and a more far-reaching description of the agent's activities as an explanatory connection? If the connection were a causal one, or somehow reducible to a causal connection, that would be an answer. This objection to the context-placing theory of rationalising explanations has been raised by Megan Fritts (2021), who argues that putting something in context, thereby making it intelligible, and explaining something are 'usually considered two different goals or activities'. I am convinced that rationalising explanations are context-placing and serve to make an agent's action intelligible in just the way that Tanney, Thompson and Mayr describe. A crucial function of rationalising explanations is that they redescribe an agent's action in such a way as to situate the agent's action into the agent's wider activities. However, I also agree with Fritts that this is not quite enough to meet Davidson's challenge. What's missing is an account of why context-placing qualifies as explaining.

8.2.3 Structural explanations

Fritts (2021) suggests that rationalising explanations are structural explanations. Fritts takes as her starting point the idea hit upon by Tanney, Thompson and Mayr that rationalising explanations explain by situating an agent's action within a wider pattern of activity the agent is engaging in which thereby makes the action expected. As Fritts puts it, 'if … our reasons for action are, at bottom, other activities in which we are involved—then human activity has a nesting-doll structure where smaller actions are constituents of larger activities' (2021: 20).

Fritts then argues that because rationalising explanations are explanations that appeal to the 'nesting-doll structure' of intentional actions, as opposed to the triggers or causes of the action to be explained, this makes them structural explanations.

Fritts endorses Stuart Shapiro's (1997) definition of a structure. Shapiro defines a structure as 'the abstract form of a system' where a system is 'a collection of objects with certain relations'. A structure is the form of a system, which is to say it is something that describes 'the interrelationships among the objects'

and ignores 'any features of them that do not affect how they relate to other objects in the system' (Shapiro 1997: 73). Explanations that appeal to the form of a system are structural explanations. According to Fritts, rationalising explanations are explanations that appeal to the fact that the action to be explained exists within a system of interrelated activities that place constraints on what activities can/should be performed.

Sally Haslanger (2016) gives a nice example of a structural explanation:

> Suppose I am playing ball with my dog. I stuff a treat into a hole in the ball and throw it for him. The ball goes over the lip of a hill and rolls down into a gully. Why did the treat end up in the gully? If we imagine the trajectory of the treat alone, from a space near my hand, through an arc in the air, then landing about an inch above the ground and moving at about that height down the hill until it stops, it would be a huge task to explain the particular events that determined each of its movements. A much easier explanation would be to point out that the treat was inserted into a ball that was thrown and rolled down the hill into the gully. In this latter explanation, we explain the behaviour of the treat by its being part of something larger whose behaviour we explain. (2016: 114)

The structural explanation for why the treat ended up in the gully has a distinctive advantage over the event-causal explanation, which is that only the latter tells us why the treat would still have landed in the gully even if Haslanger's ball-thrower had thrown the ball slightly differently. Throwing the ball slightly higher or with slightly more force would not have made a difference, as what mattered for the treat ending up in the gully is that it was inside the ball. Haslanger suggests that the structural explanation 'provides a better model for seeing how I could intervene to prevent the treat from ending up in the gully (not throw the ball in that direction, for example, or catch up with the ball and stop it from rolling)' (2016: 115). This feature of structural explanations has parallels in the rationalising explanation case—lending support to Fritts's proposal that rationalising explanations are structural.

Consider for example rationalising explanation (b): Beth is buying flour because she wants to make bread. We might be able to give an explanation of Beth's buying flour that starts with the onset of her desire to make bread, involves brain activity and muscle movements, and ends with her at the cash register paying for flour. However, that explanation would not, on its own, tell us why Beth would still have ended up buying flour even if she had, say, driven rather than walked to the shop, or walked to a different shop, or spent a day doing something else before buying flour. However, explaining Beth's buying flour by pointing out that it is part of Beth's intentional bread-making activities does provide this information. Both the event-causal explanation and the structural explanation involve the mental concept 'wanting to make bread'. The former uses this mental concept to pick out a mental event: the onset of a

desire to make bread. The latter uses this mental concept to denote the 'nesting-doll' structure of Beth's activities.

Does this account of rationalising explanations meet Davidson's challenge? The problem with Tanney's context-placing account of rationalising explanations is that it does not tell us what qualifies the connection between the agent's action and a more far-reaching description of the agent's activities as an explanatory connection. Fritts's development of the context-placing account seems to answer this question: the connection is the kind of connection we see in structural explanations—i.e. one that connects the explanandum with the form of the system of which the explanandum is part.

However, an issue with this account of rationalising explanations is whether structural explanations are really *non-causal* explanations. If you take the Davidsonian view of what a causal explanation is, then structural explanations are not causal explanations. They do not function by identifying the cause of the explanandum; therefore, they are not causal explanations. However, in the previous chapter, I sketched an alternative theory of what makes an explanation causal. I stated that causal explanations are those that provide information relevant to the manipulation of an effect. They are explanations that provide us with information about how to stop something from happening, or how to get something to happen again, or how to get it to happen in a different way (or at least information about how to make such outcomes more likely). Structural explanations seem to provide this kind of information. Haslanger's example of a structural explanation tells us how to prevent the treat from ending up in the gully, even if it does not tell us the cause of the treat's falling into the gully. If structural explanations give us information relevant to the manipulation of an effect, then they would count as causal explanations, even if they do not identify causes of the explanandum.[39]

Furthermore, the fact that rationalising explanations can provide information that would tell us how we might stop an agent from performing an action seems to be relevant to their qualifying as explanations and not mere rationalisations. A key difference between (a) and (a') is that (a') gives us information about how we could have persuaded Anna not to speak at the conference. Telling Anna that speaking at the conference would not be good for her career would not have made a difference to her action. However, telling Anna that speaking at the conference would not impress her friends might have prevented her action. This is one way to explain the difference between a reason Anna thought justified her action but which was not the reason for which she acted, and the reason she acted in light of. Nevertheless, even if structural explanations are a species of causal explanation, the theory that rationalising explanations are structural explanations shows that it is not necessary, to meet Davidson's challenge, to construe rationalising explanations as somehow identifying *the cause* of the action they explain, even if we have to acknowledge that part of

[39] Skow (2018) makes a similar argument.

what makes rationalising explanations explanatory is that they provide causal information. The position we have arrived at lies in between the causalist and non-causalist views. Rationalising explanations are causal explanations, but they are not explanations that function by identifying the cause of the action they explain. Instead they explain as structural explanations do: by identifying the wider structure of the agent's activities, of which the explanandum action is a part.

8.3 Rationalising explanations as unique causal explanations

The debate between causalists and non-causalists is a difficult debate to adjudicate on because both sides have intuitive appeal. However, there is a way to accept the non-causalist's ideas about mental concepts and how rationalising explanations explain without giving up the intuition that rationalising explanations are causal. The assumption made by both Davidsonians and non-causalists is that an explanation is causal only if it depends for its truth on the obtaining of a causal relation. Davidsonians and non-causalists both assume that there's one sort of thing causation can be; therefore, what real-world facts an explanation can answer to does not vary according to the explanatory context. This is a key part of the relational approach to causation that I have sought to challenge. The relational approach to causation says that there is one kind of causal reality true causal explanations answer to. However, as I argued in Chapter 7, my non-relational approach to causation allows us to argue that there are some causal explanations that are not made true by a pair of causally related events. Because causation is a term that can refer to processes and dynamic states of affairs as well as difference-making relations between events, there is more than one kind of causal reality that causal explanations can answer to. As I argued in Chapter 7, explanations can be causal even when they do not necessarily imply the existence of causal relations between certain particulars.

It is possible, therefore, that the peculiar features of rationalising explanations—features that set them apart from more typical event-causal explanations—are not barriers to thinking of these explanations as causal. It is possible that rationalising explanations could be causal even though the mental concepts cited in the rationalising explanation do not designate causally efficacious items. Rationalising explanations could be the kind of causal explanation that answers to the non-relational aspect of causal reality.

This thesis, that rationalising explanations are causal explanations that are made true by the non-relational aspect of causal reality, is attractive for at least two reasons. First, it allows us to save the intuition that explaining someone's actions in terms of their beliefs and desires is to give causal information, while at the same time accepting that the mental concepts appealed to in rationalising explanations do not refer to items that stand to the action explained as cause to effect. In other words, the thesis that rationalising explanations are

causal, but made true by the non-relational aspect of causal reality, allows us to acknowledge what's intuitive about both the Davidsonian and the non-causalist views. We can agree with non-causalists like Anscombe and Tanney that when we explain an agent's action by giving their reasons we are not identifying the trigger of their action, or that which made the difference to their action occurring. Instead, we are explaining why a person (a substance) engaged in a particular activity. The action that comes about when the agent completes her activity may well have a difference-making cause—but that is not what we are interested in when we give a rationalising explanation. What's more, whether it has a difference-making cause is often not relevant to the truth of the rationalising explanation.

Second, there are similarities between rationalising explanations on the one hand and process-citing and disposition-citing explanations on the other, which lends support to the idea these three kinds of explanation belong in the same general category. Some rationalising explanations appear to be very similar to causal explanations that cite the continuous operation of causal processes. Causal explanations that cite the continuous operation of causal processes are roughly of the form: some effect occurred or is occurring, or obtained or obtains, because substance S is or was engaging in causal process P. Thompson's (2008) 'naïve action explanations' have this form, as they explain why an agent engaged in some activity in terms of something else the agent is doing. Other rationalising explanations have the form of stative explanations. If, as seems plausible, mental states like desiring, believing and knowing are dispositions, then this would make those stative rationalising explanations disposition-citing explanations. Hyman (2015: 103–132) and Mayr (2011: 295) also propose that rationalising explanations that cite mental states of the agent are disposition-citing explanations.

However, there are two ways in which rationalising explanations are unique. First, if mental states like desiring, believing and knowing are dispositions, they are not ordinary dispositions. Most dispositions are dispositions to engage in or undergo a certain specific activity or process. In contrast, having a desire to do something or achieve something (for example) disposes one to undertake whatever activities are deemed, by the agent, to be acceptably good means of achieving what one wants; to deliberately refrain from acting should *that* turn out to be an acceptably good means of achieving what one wants; to feel happy or pleased if one's desire gets satisfied or disappointed if it is frustrated; and to use one's desire as a premise in practical deliberation about what to do. Desires are not dispositions to do any one specific thing (or even any two specific things)—they are rather dispositions for one's activities to instantiate a certain pattern or goal-directedness, which is made sense of by the content of the desire. Similar claims can be made about other mental concepts. As Ryle suggests, it would be wrong to think, just because the verbs 'know' and 'believe' are 'ordinarily used dispositionally', that 'there must therefore exist one-pattern intellectual processes in which these cognitive dispositions are actualised'

(1949: 44). Rather, states of believing and states of knowing, if they are dispositions at all, are 'dispositions the exercise of which are indefinitely heterogeneous' (1949: 44). So, while there are some similarities between rationalising explanations on the one hand and process-citing and disposition-citing explanations on the other, it is important not to forget that rationalising explanations are unique: they are very variable in form and, even if we suppose that the mental states cited in rationalising explanations are dispositions, they are not, by any means, ordinary dispositions.

Second, rationalising explanations do not exactly provide us with information about how to stop something from happening, or how to get something to happen again, or how to get it to happen in a different way (or at least information about how to make such outcomes more likely). When you learn that some agent's activity is a manifestation of her desire or an output of her rational capabilities, you learn that you might be able to alter her activity by altering what she believes about the world, or by changing her desires, perhaps by changing her environment but more usually by reasoning with her, talking to her or persuading her. However, learning this information only makes it the case that you *might* be able to alter the agent's activity. This is because reasoning with an agent in an attempt to get them to φ, for example, does not guarantee that the agent will φ—it does not even ensure that it is more likely that the agent will φ. This is because the agent can ignore you, or remain unconvinced, or even just act against her better judgement. In short, rationalising explanations do not seem to be the sort of explanations that provide us with information about how to stop something from happening, or how to get something to happen again, or how to get it to happen in a different way, or even how to make such outcomes more likely. They seem only to provide information about how we *might* stop something from happening, or get something to happen again, get it to happen in a different way, or make such outcomes more likely.

In this chapter, I have argued that explanations of intentional action that cite the agent's reasons for acting are the kind of causal explanation that are not made true by causally related events. The most important consideration favouring this view is that it saves two strong intuitions: (a) that reason-giving explanations are causal, and (b) that the mental states cited in reason-giving explanations do not denote items that stand in causal relations to the actions they explain. This view has important consequences for how we ought to think about the nature of intentional action. As mentioned, it is commonly held that we can achieve an adequate account of what it is to act intentionally by examining the distinctive sort of explanation with which intentional actions are associated. If rationalising explanations are causal explanations that do not designate mental items that stand to the action explained as cause to effect but instead answer to non-relational causal reality, then the case for thinking intentional actions are distinguished from non-intentional actions by their mental causes is significantly weakened. However, without getting clearer on exactly what facts about dynamic states of affairs rationalising explanations could plausibly be said to answer to, it

is difficult to offer a positive account of what the distinguishing mark of intentional action is. In the next chapter, I will present a view on intentional action that grants that some rationalising explanations are disposition-citing explanations and others are structural explanations but which also respects the two ways in which rationalising explanations are unique.

References[40]

Anscombe, G E M 2000 *Intention*. Cambridge, MA: Harvard University Press. Originally published in England in 1957 by Basil Blackwell.

Child, W 1994 *Causality, interpretation, and the mind*. New York: Oxford University Press.

Curry, D 2018 Beliefs as inner causes: The (lack of) evidence. *Philosophical Psychology*, 31(6): 850–877. DOI: https://doi.org/10.1080/09515089.2018.1452197

Dancy, J 2000 *Practical reality*. New York: Oxford University Press.

Davidson, D 1963 Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700. DOI: https://doi.org/10.2307/2023177. Reprinted in Davidson 2001a pp. 3–20.

Davidson, D 1997 Indeterminism and antirealism, In: Kulp, C B *Realism/antirealism and epistemology*. Lanham, MD: Rowman & Littlefield. pp. 109–122. Reprinted in Davidson 2001b pp. 69–84.

Davidson, D 2001a *Essays on actions and events*. 2nd Edition. Oxford: Clarendon Press.

Davidson, D 2001b *Subjective, intersubjective, objective*. Oxford: Clarendon Press.

DiYanni, C and Kelemen, D 2005 Using a bad tool with good intention: How preschoolers weigh physical and intentional cues when learning about artifacts. *Cognition*, 97: 327–335. DOI: https://doi.org/10.1016/j.jecp.2008.05.002

Fodor, J 1987 *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fritts, M 2021 Reasons explanations (of actions) as structural explanations. *Synthese*, 199(5–6): 12683–12704. DOI: https://doi.org/10.1007/s11229-021-03349-4

[40] Author note: some references to Davidson are formatted (1963/2001a). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001a collection of his essays, with the page numbers relating to that volume. Similarly, some references to Davidson are formatted (1997/2001b) which indicates the initial date of publication (in this case 1997) but references the paper as it appears in the 2001b collection of Davidson's essays, with the page numbers relating to that volume.

Hart, H L and Honoré, A M 1985 *Causation in the law*. 2nd ed. Oxford: Oxford University Press.

Haslanger, S 2016 What is a (social) structural explanation? *Philosophical Studies*, 173(1): 113–130. DOI: https://doi.org/10.1007/s11098-014-0434-5

Hatano, G and Inagaki, K 1994 Young children's naive theory of biology. *Cognition*, 50(1–3): 171–188. DOI: https://doi.org/10.1016/0010-0277(94)90027-2

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Löhrer, G and Sehon, S 2016 The Davidsonian challenge to the non-causalist. *American Philosophical Quarterly*, 53(1): 85–96.

Lombrozo, T and Carey, S 2006 Functional explanation and the function of explanation. *Cognition*, 99(2): 167–204. DOI: https://doi.org/10.1016/j.cognition.2004.12.009

Malle, B 1999 How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1): 23–48. DOI: https://doi.org/10.1207/s15327957pspr0301_2

Malle, B 2004 *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B, Knobe, J and Nelson, S 2007 Actor–observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, 9(4): 491–514.

Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.

Monroe, A and Malle, B 2017 Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, 146(1): 123–133. DOI: https://doi.org/10.1037/xge0000234

Ryle, G 1949 *The concept of mind*. London: Hutchinson's University Library.

Schult, C and Wellman, H 1997 Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition*, 62(3): 291–324. DOI: https://doi.org/10.1002/icd.548

Sehon, S 2005 *Teleological realism: Mind, agency, and explanation*. Cambridge, MA: Bradford Book/MIT Press.

Sehon, S 2007 Goal-directed action and teleological explanation. In: Campbell, J K, O'Rourke, M and Silverstein, H S *Causation and explanation*. Cambridge, MA: MIT Press. pp. 155–170.

Sehon, S 2010 Teleological explanation. In: O'Connor, T and Sandis, C *A companion to the philosophy of action*. Oxford: Wiley-Blackwell. pp. 121–128.

Shapiro, S 1997 *Philosophy of mathematics: Structure and ontology*. New York: Oxford University Press.

Skow, B 2018 *Causation, explanation, and the metaphysics of aspect*. Oxford: Oxford University Press.

Tanney, J 2009 Reasons as non-causal, context-placing explanations. In: Sandis, C *New essays on the explanation of action*. Basingstoke: Palgrave Macmillan, pp. 94–111.

Tanney, J 2013 Ryle's conceptual cartography. In: Reck, E H *The historical turn in analytic philosophy*. Basingstoke: Palgrave Macmillan.
Thompson, M 2008 *Life and action: Elementary structures of practice and practical thought*. Cambridge, MA: Harvard University Press.

CHAPTER 9

A New Theory of Intentional Action

One of the main aims of this book is to explain how physicalism, causal theories of intentional action and a relational approach to causation are linked. I argued in the first half of this book that these three theoretical positions are mutually supporting and form what I called the physicalist triad. I argued that we have good reason to reject the physicalist triad because the picture of human agency the triad entails is inadequate. The chief failing of the physicalist/event-causal account of agency is that it eliminates the agent from the causality of her action, which contradicts an essential part of our concept of agency—that the agent herself brings about changes. This is known as the disappearing agent objection. Agent-causal accounts of agency avoid the disappearing agent objection as they construe agency as a kind of causation where the agent exercises causal power and this exercise of causal power cannot be reduced to causation by an event involving the agent. However, I argued in Chapter 5 that agent-causal accounts face a number of issues because the metaphysical assumptions about causation they rely on are not sufficiently distinct from the relational approach to causation.

In the second half of this book, I started to navigate a path out of the physicalist triad. In Chapter 6 I proposed a non-relational theory of causation. According to this theory, causation is not always a relation but can be a process that substances engage in. Chapters 7 and 8 were concerned with explaining how this non-relational theory of causation allows us to challenge the standard causal theory of action explanation. The non-relational theory of causation allows us to think of rationalising explanations as providing causal information even though the concepts employed in such explanations do not designate causes of the actions they explain. This has consequences for how we ought to understand what intentional action is. The task of the present chapter is to make good on my promise that a non-relational theory of causation, and the

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 175–195. York: White Rose University Press. DOI: https://doi.org/10.22599/White.j.

ontology that permits it, supports an alternative view of intentional action. I propose that to act intentionally is to engage in a process, and as such is to exercise a power—but a power of a special sort. The power to act intentionally is a power to structure one's own activities so that they demonstrate a pattern—a pattern that is only revealed by attributing mental states to the agent.

9.1 A neo-Aristotelian theory of agency

Before turning my attention to *intentional* action, it is necessary to say something about what action in general is. Part of the task of philosophy of action is to explain what agency is, or what it is to act. Like other agent-causal accounts, I propose that we understand agency in terms of substance causation. Like other agent-causationists, I believe that agency is a kind of causation where the agent, who is taken to be a substance not an event, exercises causal power and this exercise of causal power cannot be reduced to causation by an event involving the agent. However, my account of what substance causation is differs from standard agent-causal accounts. In Chapter 6, I outlined a distinctive non-relational understanding of substance causation that made use of a novel process ontology. I said that processes are universals and can be described as ways for substances to be changing, to be effecting change or to be resisting change. Processes that are (to some degree) ways for substances to be effecting change are species of causation. These mostly active processes I will call *activities*. What it is for a substance to be causing something is for there to be an *activity* that the substance is engaging in. A substance engaging in an activity is an agent, and the event that results once the substance has completed the activity it has been engaging in is an action. Actions are thus events of a special kind: they are events that are instances of activities.

Importantly, agents are not causally related to their actions. Individual actions are events that come into existence when an agent engages in an activity and then completes that activity. So understood, actions are 'produced by' or 'brought into being by' agents, but the sense of production here is ontological not causal. This metaphysics of action distinguishes my account from standard agent-causal accounts, which take substance causation to be a relation between a substance and an event. It also helps us see why the causality of action is something that essentially involves the agent (and thereby avoids the disappearing agent problem). On my theory, the causation exemplified by actions is the activity the agent engages in; it is something that goes on, but only insofar as it is engaged in by an agent. Furthermore, the dynamic state of affairs that is an activity going on is something that is partially constituted by the agent. A dynamic state of affairs is, as I proposed in Chapter 6, a complex entity composed of a substance and a process. So, if we take the causality of action to be a dynamic state of affairs, then the agent herself *partially constitutes* the causality

of action—she cannot, therefore, be merely the arena within which the causation of her action takes place.

Hornsby has described views like mine as 'neo-Aristotelian':

> Neo-Aristotelians do not treat cause as everywhere a relation—neither as a relation between two events, nor between two objects, nor between an object and an event … They take an object's powers to tell us what kinds of processes the object can engage in, so that they connect our understanding of causality with our recognition of the display of the potentialities of things by the things having those potentialities. Thus they defend a metaphysics in which a substance ontology belongs, and to which such notions as powers, capacities, liabilities are central … Causality, then, is present in the world inasmuch as something is actually exercising its powers, perhaps affecting something else in doing so. (2015: 131–132)

The theory I have just proposed tells us what sort of entity an action is (an event, i.e. an instance of activity). My theory also tells us what sort of entity the exercise of power is: the exercise of power by a substance is a dynamic state of affairs, i.e. a substance's engaging in a process. However, providing a *metaphysics of action* is not all that is required for a complete and adequate theory of *agency*. It takes more to provide an adequate theory of agency than simply to describe the ontological structure of the worldly entities that are picked out by the concepts of *action, agent* and *activity*. To provide a complete theory of agency, one must consider the concept of agency and provide some sort of dissection of this concept.

I believe there are two distinctions crucial to our concept of agency: the distinction between activity and passivity, and the distinction between one-way and two-way powers. Agency cannot be identified with either the exercise of active power or with the exercise of two-way power. Instead, *both* concepts are key to understanding agency. The agency concept has something to do with the idea of agents as things that bring about change. John Hyman suggests that 'to act is to intervene, to make a difference, to make something happen, to cause some kind of change' (2015: 33). Agents *cause* change and should be contrasted with patients, who *undergo* or *suffer* change (Hyman 2015: 34). On this understanding of agency, plants, animals and inanimate objects can be agents. They are agents whenever they cause something to happen. I agree with Hyman that the concept of *agent* is kindred with *causation, production* and *activity*, so the notion of active power is essential to understanding what agency is. It might sound strange to say that inanimate objects can be agents but denying that inanimate objects can act is at odds with the language we use to report actions. We typically report actions by means of causative verbs like 'melt', 'burn' and 'pump'. But we say things like 'the acid melted the beaker', 'the poker burnt the cloth' and

'his heart pumped blood' just as readily as we say 'the cook melted the butter', 'the criminal burnt the evidence' and 'the man pumped the water'. As Hyman (2015: 30–31) has argued, it is implausible to think that these verbs have different meanings when they are used to report what inanimate things have done and when they are used to report what human beings have done.

Even though I think it literally true that inanimate objects can be agents, and they are agents when they exercise active power, there is more to the concept of agency than activity. Agency and activity are not synonyms. It seems to me that one has not really mastered the concept of agency until one has recognised the difference between things that lie there until something else comes along and prods them into action, and things that, sometimes with effort, move themselves about. It seems to be an essential part of our concept of agency that acting must involve a very minimal kind of autonomy.

Agency is connected to the idea of being able to move oneself. It contrasts with what Aristotle called 'moved-movement'. Therefore there is an important difference between the agency of inanimate objects and the agency of animals and human beings—and understanding this difference is essential to understanding the agency concept. This is because, as well as being kindred with concepts like *causation, agency* is associated with ethical concepts like *responsibility* and *blameworthiness*. As Hyman (2015) puts it, some instantiations of agency have an 'ethical dimension' as well as a 'physical dimension'. It is of great ethical significance that some actions are up to the agent whereas others are not up to the agent. There is an important moral difference between pushing someone over when you could have refrained from doing so and pushing someone over because someone else pushed you into them. This distinction has something to do with agency, and I think the terms 'settling', 'self-movement', 'up-to-us-ness' and 'origination' are all different ways philosophers have attempted to describe this crucial contrast. I think the best way to understand this contrast is using the concept of a two-way power.

I endorse Kim Frost's definition of a two-way power as one that has 'two fundamental, mutually exclusive kinds of exercise', whereas a one-way power has only one fundamental kind of exercise (2013: 612). The easiest way to spell out this idea is by means of an example. In the right circumstances my power to sing is two-way. What this means is that, if I do end up singing, I am manifesting my two-way power, but if I end up *not* singing (which might involve actively doing something else, but might not—it might involve continuing an activity already in progress, or letting something happen to me), I am *also* manifesting my two-way power. Thus, my power to sing, because it is two-way, is sometimes manifested by singing, and sometimes manifested by *not* singing. The power has two mutually exclusive kinds of exercise, which I will call positive and negative, and only one of these (the positive) is the activity the power is specified as a power to do.

In the case of one-way powers, when the conditions are right for the manifestation of a one-way power, the activity the power is a power to do will be

engaged in, whereas in the case of two-way powers, when the conditions are right for the *positive* manifestation of a two-way power, the two-way power may *not* be exercised positively—it may be exercised negatively—and thus the activity the power is a power to do may not be engaged in. It is important to note that, while one-way powers can be distinguished into those that are active and those that are passive, the active–passive distinction does not have application in the case of two-way powers. This is because two-way powers are powers to act *or refrain*, so they are all powers to be active in a certain way, or not (which might be to be active in a different way, or might be to be passive).

Steward (2013a) finds the conception of two-way powers as powers with two distinct fundamental kinds of manifestation problematic. For Steward, a power to φ is two-way just in case the agent who possesses the power to φ *also* possesses the power not to exercise their power to φ (2013a: 691). Steward argues that a conception of two-way powers like mine (and Frost's) has counterintuitive consequences (2013a: 691). As Steward notes, it seems to entail that in not singing right now while I'm working on this chapter, I am exercising my power to sing, albeit negatively. I accept that it is counterintuitive to think that, in not singing right now, I am exercising my power to sing. It is more intuitive to think that my power to sing is dormant while I am working on this chapter: it is not being exercised at all. I thus acknowledge that not *every* case where an agent does not φ counts as a negative exercise of a two-way power to φ; not every case of not doing something is a case of *refraining* from doing it. However, I think a conception of two-way powers as powers with two mutually exclusive kinds of exercise is compatible with the fact that not every case of not doing something is a case of *refraining* from doing it.

As long as one can say something about how to distinguish cases where a two-way power to act is exercised negatively from cases where the power to act is just not exercised at all, then one is permitted to claim that there is more to exercising a two-way power to φ negatively than simply not φing. I doubt that there is a completely general way to distinguish cases where an agent exercises her two-way power to φ negatively from cases where an agent's not φing does not count as a negative exercise of her two-way power to φ. This is because what it takes for some instance of not acting in a certain way to count as refraining from acting in that way might depend on the type of action in question. For example, the fact that I am consciously aware of my cup of coffee might be sufficient for my not reaching for the cup to count as a negative exercise of two-way power to reach for it. But, for my not singing right now to count as a negative exercise of two-way power to sing, I may need indexical knowledge that the circumstances I am in are circumstances in which I could (or should) be singing. In all cases of refrainment, I think some sort of awareness of what one could be doing is required, but precisely what sort of awareness is required differs depending on the type of action in question.

Maria Alvarez (2013) argues that, most of the time when human beings exercise their agential powers, the power they are exercising is a two-way power.

This falls in line with the intuition that most of the time human agency is 'self-movement' and involves at least a minimal kind of autonomy. This minimal autonomy consists in our activities being up to us, in the sense that our power to perform these activities is two-way.

One challenge facing any theory of agency that appeals to two-way powers is whether this entails that agency is incompatible with determinism, the doctrine that every event is completely causally determined by prior events and conditions together with the laws of nature. Helen Steward (2012) argues that it does, and so much the worse for determinism. In other words, because agency must be understood as a two-way power, if this entails that agency is incompatible with determinism, then determinism must be false as it is undeniable that agency exists. Possibly, if one were convinced of the truth of determinism one could then argue for the non-existence of agency, just as hard incompatibilists argue for the non-existence of free will. However, denying the existence of agency seems a very high price to pay. A more plausible strategy for those convinced of the truth of determinism is to argue that possessing two-way powers is compatible with determinism.

A number of compatibilists have argued that determinism is compatible with possessing the ability to do otherwise. Some of these compatibilist arguments could be used to show that possessing two-way powers is compatible with determinism. This is because a necessary condition for having a two-way power to φ at a time t is to be able both to φ and not φ at t (Alvarez 2013: 108).[41] The kind of compatibilist argument that could be used to show that possessing two-way powers is compatible with determinism are those that analyse the ability to do otherwise *modally*, i.e. the agent is able to do otherwise just in case it is possible for the agent to do otherwise.[42] Compatibilist arguments that analyse the ability to do otherwise *conditionally*, i.e. the agent is able to do otherwise just in case they would have done otherwise had they tried to (or intended to, or chosen to), would not work. This is because, on a conditional analysis of the ability to do otherwise, an agent cannot possess the ability not to φ whenever she is able to φ (the necessary condition for possessing a two-way power to φ). Possessing the ability not to φ is conditional on what the agent tries/intends/chooses: if they try to φ at t, then they do not possess the ability not to φ at t. Compatibilist arguments that analyse the ability to do otherwise *modally* could be used to defend the idea that two-way powers are compatible with determinism. These

[41] Another necessary condition for having a two-way power to φ at a time t is to have the opportunity both to φ and not to φ at t (Alvarez 2013: 108). If agent A has the ability to φ, then she has the right attributes for φing and knows how to φ (for example, A only has the ability to wave her arms if she has arms and knows how to wave them). If A has the opportunity to φ, then there is nothing preventing her from φing (for example, she is not tied up or injured). See also: Kenny (1975: 33).

[42] Berofsky (2011), Campbell (2005), Kapitan (2011) and List (2014).

arguments turn on the idea that there can be more than one meaning of 'possible'. This allows one to argue that, even if determinism entails that only a φing action at t is *physically* possible given prior events and conditions and the laws of nature, it may still be possible in another sense for the agent to not φ at t. For example, it could still be *agentially* possible for the agent to not φ at t.

I will not adjudicate on the question of whether possessing a two-way power is compatible with determinism here. The fact that it is possible to argue that agency understood as a two-way power is both compatible with determinism and incompatible with determinism suggests that perhaps agency cannot settle the question of whether determinism is true or not.

Still, recognising that human agency is often the exercise of a two-way power has several advantages.

First, it can explain why there is no intentional action in deviant causal chain cases. As mentioned in Chapter 4, deviant causal chain cases are a well-known problem for event-causal analyses of intentional action, i.e. analyses that attempt to reduce intentional action to causation of bodily movements by appropriate mental states and/or events. The most famous deviant causal chain case is Davidson's (1973/2001: 79) example of a climber whose desire to rid himself of the weight of carrying another man and belief that he could do so by loosening his hold cause him to become so nervous that he lets go unintentionally.

For the event-causal theorist there is no intentional action in this case because the causal chain does not follow the sort of causal path that counts as 'the "right" way in which beliefs and desires must yield behaviour for genuine intentional action to occur' (Bishop 1989: 135), the 'right way' being '…', where the '…' has to be filled in without reference to intentional action. The success of this explanation depends on how the '…' is filled in and, as we saw in Chapter 4, no account of how the '…' ought to be filled in has been completely counterexample-free.

An alternative explanation is made available if we assume that exercising a two-way power is necessary for intentional action. If possessing and exercising a two-way power is a necessary condition for acting intentionally, then there is no intentional action in deviant causal chain cases because the agent's reasons or intentions or mental states rob the agent of the relevant two-way power, most probably by robbing the agent of the opportunity to both φ and not φ. For example, in Davidson's example, the climber's nervousness robs the climber of the opportunity not to let go of the rope. Just as extreme grief can render a person incapable of not crying out, the climber's control over his body has been hijacked by the conditions responsible for his nervous state. It is no longer up to him whether he lets go or not.

We can also now explain why some heteromesial cases are such that intentional action is blocked, and others do not block intentional action: not every heteromesial case is such that the agent is stripped of either the ability to φ or not φ or the opportunity to φ or not φ. When Amy is using her device just to keep my neural systems in working order, she has not robbed me of the ability

or opportunity to not make tea, which is why I am still exercising agency in that example, whereas where she uses her machine to control the movements of my body she has robbed me of the opportunity not to make tea.

Another advantage of explaining agency in terms of two-way powers is that we can now explain how agency can be demonstrated in passivity as well as in activity. When one's agential power is two-way, one can demonstrate this power by *not* performing the action one's agential power is a power to do. For example, in cases of intentional refrainment, e.g. where I let my plant die by not watering it or allow a telephone to continue ringing by not answering it, the putative agent exercises a two-way power to act negatively. In failing to water my plant, I do not *actively* cause the death of the plant. Substances in the vicinity that might have actively caused the death of the plant probably include parts of the plant itself (e.g. the plant's chloroplasts may have actively caused the death of the plant by using up what water was stored in the plant, thereby causing the plant to wilt, which in turn prevented the plant from capturing light etc.). In this case, I demonstrate agency by letting the active powers of *other* substances manifest themselves, rather than by exercising any active powers myself. In this case, I possess a two-way power to water the plant and I exercise my power to water the plant *negatively*. In Hyman's example of a child allowing themselves to be picked up, the child is demonstrating agency because the child is manifesting her two-way power to resist being picked up (e.g. by pushing away the parent) negatively. So, even though the child is, so to speak, not doing anything but rather letting something happen to her, she is demonstrating an agential power.

I also think that using *both* the distinction between active and passive powers and the distinction between one-way and two-way powers to explain what agency is has a distinctive advantage. The question 'what marks the difference between things that one does, and things that befall one?' is a complicated question. It is complicated because there are lots of different distinctions that have a bearing on it: the distinction between causing change and suffering change; the distinction between automatic behaviours and intentional ones; the distinction between moving oneself and being moved to move by something else. Appealing to both active and passive powers and one-way and two-way powers can help clarify this question. Agency does not reduce to the exercise of active power, because some substances can manifest their agency by remaining passive, and therefore by not engaging in activity. Neither does agency reduce to the exercise of two-way power, because not all substances that cause things to happen do so by exercising two-way powers, but all substances that cause things to happen are agents. My view is that agency is a complex, highly abstract concept that incorporates both distinctions. Some substances' agential powers are one-way; these substances manifest their agency when they are active but not when they are passive. For these substances, exercising their agential power is to engage in an activity. Other substances' agential powers are two-way; these substances manifest their agency when they are active but also

sometimes when they are passive. For these substances, in *some* cases exercising their agential power is to engage in an activity, but in other cases exercising their agential power is to allow other substances to act upon them.

Understanding agency using both the active–passive distinction and the distinction between one-way and two-way powers also has the advantage of giving us more conceptual resources for discussing some of the tricky cases discussed in Chapter 4, including reflexes, sub-intentional action, and spontaneous expressions of emotion.

Most of us would agree that reflexes, like blinking or blushing or sneezing or the knee-jerk reflex, are not intentional. Opinions are more divided on the question of whether reflexes are genuine actions. We call them 'reflex actions' and they are things that we do. However, they are not activities over which we have any kind of control. It is not up to me whether or not I blink when an object touches my eye; when a doctor hits my patella tendon with a reflex hammer I cannot but move my leg. For this reason, it seems wrong to attribute reflex actions to the person. Instead, reflex actions are more properly attributable to sub-personal systems. They are controlled by neurons in the spinal column and lower parts of the brain. When we perform reflex actions, we seem to be 'moved-movers'. We are moved to move by sub-personal systems. When we perform reflex actions, we are like ASIMO: our movements are strictly governed by our component parts.

Using the two distinctions that I believe are crucial to understanding agency, we can explain why reflex actions are called actions and described as things that we do even though it would be wrong to think of them as genuine demonstrations of human agency. Reflex actions can sometimes count as exercises of causal power. Suppose I kicked over and broke a vase as a result of stimulation of the knee-jerk reflex. In this case, I caused the vase to break and so I exercised a causal power. I was active rather than passive with respect to the breaking of the vase (though I was passive with respect to moving my leg—*I* didn't get my leg to move; the doctor and my own sub-personal systems did), so in a sense I was the agent of the vase's breaking. However, the power I exercised here was one-way and not two-way. I could not have refrained. Human beings are the kind of creatures whose movements are often up to them, hence we are the kind of creatures whose agential powers are two-way. Given this, reflex actions are not genuine demonstrations of human agency because they are not exercises of two-way power.

What about sub-intentional actions and spontaneous expressions of emotion? In Chapter 4, I described these as examples of agency that were nevertheless not intentional. The reasons I outlined for counting these as examples of agency were (a) because they are attributable to the person and not to another agent or sub-personal system; (b) because it is natural to speak of the person moving their body in cases of sub-intentional action and spontaneous expressions of emotion—in other words, they seem to be examples of self-movement; and (c) sub-intentional actions and spontaneous expressions of emotion seem

to be behaviours over which we are in control. I can now add that these examples count as demonstrations of agency because they are exercises of two-way power. When I absent-mindedly fiddle or tap my feet to music, I have the ability and opportunity not to engage in that behaviour, and that is what my control over the activity consists in. When I spontaneously embrace a loved one or laugh at a joke, again, I have the ability and opportunity not to, which is why it is true to say that engaging in these activities is up to me.

One could object to the idea that sub-intentional actions and spontaneous expressions of emotion are exercises of two-way powers. Alvarez (2013: 113) lists spontaneous expressions of emotion as actions that 'we cannot generally avoid doing' and thus as counterexamples to the thesis that human agency involves the exercise of two-way power. One could perhaps say the same about sub-intentional actions—they are activities we cannot generally avoid doing. However, Alvarez offers a good response. She suggests that spontaneous expressions of emotion (and presumably sub-intentional actions too) lie on a continuum that ranges from out of our control and attributable to sub-personal systems, to under our control and attributable to us. Another way of putting this point is to say that *some* spontaneous expressions of emotion are really out of our control and attributable to sub-personal systems, whereas others are within our control and attributable to us as persons, and some fall in between these two extremes. Alvarez further suggests that these activities will seem closer to one or the other end of this continuum to the extent that we are aware of our doing them. The more aware we are, the more able we are to control the activity. Alvarez then argues that whether an activity falls towards the 'controlled by sub-personal systems' end of the continuum or towards the 'controlled by us' end of the continuum depends on 'the extent to which we determine when they happen, suppress them if we choose … that is, to the extent to which doing them involves exercising a two-way causal power to move' (2013: 114).

I agree with many of Alvarez's suggestions. I agree that spontaneous expressions of emotion, and sub-intentional actions, fall onto a continuum between attributable to sub-personal systems and demonstrations of our own agency. However, I disagree with Alvarez's suggestion that doing something can be an exercise of two-way power to a greater or lesser extent. Whether or not an activity is the exercise of a two-way power seems to me to be a binary property, not something that can come in degrees. Nevertheless, I still think that the concept of two-way powers can be helpful in this case. Most of the things that we do necessitate performing a number of sub-activities. For example, to tap my foot I need to contract certain muscles in my leg. Depending on how strongly I contract these muscles I can vary how vigorously I tap my foot. Similarly, to laugh I might contract my diaphragm as well as muscles in my face and abdomen, and I can control the quality of my laughter by controlling these various contractions. There is variability in how many of these sub-activities are exercises of two-way power. Sometimes they all are. If I'm

paying particularly close attention, or if I am very skilled, I can control not only whether or not I tap my foot but also the exact manner in which I do so.[43] Sometimes, only the macro-activity is an exercise of two-way power. In this case, it could be up to me whether or not I tap my foot but not up to me exactly how I do this. (Unskilled movements are often like this.) I also think that, sometimes, the macro-activity might not be under our control but the detail might be. That is, sometimes it might not be up to me whether or not I tap my foot or laugh but it *is* up to me exactly how I do it. My suggestion is that the greater the number of sub-activities that are exercises of two-way power, the more inclined we are to say that the macro-activity is attributable to the person and not to sub-personal systems.

9.2 Intentional action

I now turn my attention to the nature of intentional action. The causal theory of action maintains that intentional actions are events. On this point, I agree. Most versions of the causal theory of action maintain that at least basic intentional actions are bodily movements. For example, the action of raising my arm is one and the same event as my arm's rising (Davidson 1987: 37). On this point, I also agree. However, this is not yet a complete answer to the question of what intentional actions are, as not all bodily movements are intentional. To complete the story, the causal theory of action maintains that events count as intentional actions when and only when they are caused, in the right way, by mental states of the agent that also rationalise the action.[44]

Much of what has been presented in Chapters 6, 7 and 8 points to the conclusion that construing intentional action as events caused to happen by mental antecedents is not the right way to understand intentionality. I propose an

[43] I do not think that attention and awareness is always what makes the difference here. For example, professional ballet dancers can control muscles in their feet that non-dancers would not be able to control. Professional dancers are therefore able to complete a wider array of very precise movements with their feet, which are necessary for being able to dance *en pointe*, for example. I would say that a professional ballet dancer can control the exact manner of her foot movements when dancing *en pointe*—that each of these finer movements was up to her—even though, while she is dancing, it is very unlikely that she is paying attention to them; she is much more likely to be thinking about what she is trying to express through her dancing. It seems to me that many highly skilled movements are like this: many of the sub-activities are up to the agent, but the agent does not need to attend to them to execute them with control.

[44] See Bishop (1989: 40–44), Davidson (1963/2001: 3–21; 1971/2001: 43–63), Mele (2003) and Smith (2012).

alternative view of intentional action. To act intentionally is to engage in a process, and as such is to exercise a power—but a power of a special sort. The power to act intentionally is a power to structure one's own activities so that they demonstrate a pattern—a pattern that is only revealed by attributing mental states to the agent. So, when an agent acts intentionally, they engage in the process of causation. The process they engage in counts as *mental* causation in virtue of the fact that the agent is manifesting a special power to organise their activities so that they instantiate a certain structure, a structure that is made comprehensible by the agent's mental states. This account builds on an account offered by Erasmus Mayr (2011).

9.2.1 Mayr's theory of intentional action

Mayr (2011) offers a theory of intentional action that takes seriously the idea that intentional action is the manifestation of a special sort of power. According to Mayr, 'intentional behaviour displays a certain characteristic structure of "purposefulness"' (2011: 271). Mayr proposes that to act for a reason is for one's behaviour to display a particular kind of structure, i.e. 'the characteristic structure of taking something as one's "standard of success and failure", or "of correctness and incorrectness"' (2011: 271). Mayr takes this proposal to be supported by the fact that, when searching for a rationalising explanation of someone's action, the facts we consider relevant are facts about whether the agent's behaviour, feelings and reasoning display—or would display—a certain pattern. For example, when we wonder if Beth is buying flour because she wants to make bread, we seek to find out things like "will Beth also buy yeast?", "if Beth got home and found out her bread tin was missing, would she feel disappointed?" and "would Beth make use of her desire to make bread in a practical deliberation?" For Mayr these facts do not merely constitute the epistemic criteria for determining what reason an agent acted in light of, they are also the facts that *make it the case* that an agent acted for a specific reason. There's nothing more to acting for a reason than for this welter of facts concerning the agent's actual or hypothetical behaviour and thinking to obtain.

What are the facts the obtaining of which makes it the case that an agent acted for a specific reason? According to Mayr's theory, there are three sorts:

1. Facts concerning the teleological structure or 'plasticity'[45] of the agent's actual or hypothetical behaviour. Mayr claims that, when an agent has a certain goal, they will 'react sensitively to changes in the environment which threaten the attainment of that goal or make it otherwise necessary to adopt different means for attaining his goal' (2011: 271)—or

[45] Mayr takes 'plasticity' to be an alternative term, used by Woodfield (1976), for this pattern in an agent's activity.

would if such environmental changes occurred. Agents with a goal will take 'corrective measures' and perform actions 'conducive to overcoming obstacles' should such mistakes or obstacles occur (2011: 271). These 'corrective movements' indicate to an observer that the agent has a 'standard by which—at least implicitly—he assesses his behaviour and considers himself—in cases of non-conformity of his behaviour to this standard—to have "made a mistake"' (2011: 273). When an agent does not encounter any obstacles or make any mistakes, the agent's actions may not display plasticity. Mayr insists that, in this case, 'our ascriptions of aims rely on our confidence that certain counterfactual conditionals about what the agent would do if obstacles arose are true, and that the hypothetical behaviour he would display would have an adequate teleological structure' (2011: 274). In other words, the plasticity of hypothetical as well as actual behaviour is important.

2. Facts concerning the agent's actual and hypothetical success and failure feelings. Achieving one's aim is often accompanied by feelings of satisfaction or joy, and failing to achieve one's aim is associated with feelings of disappointment or frustration. For Mayr, what occurrences trigger (or would trigger) feelings of satisfaction or disappointment are important for determining what the agent is aiming at, or what the agent considers to be a success and what he considers to be a failure. Of course, success is not always accompanied by feelings of joy, and failure is not always accompanied by feelings of frustration. For example, when one achieves something one considers a necessary evil, one may feel bitter and unhappy upon achieving it. In such cases, Mayr thinks that 'the only success feeling of the agent may be a half-hearted or even bitter feeling of "having done it" or "being finished"' (2011: 277).
3. Facts concerning whether the agent makes use of their purported aim as a premise in the practical deliberation leading to the action, or at least would if practical deliberation were called for. According to Mayr, when an agent is guided by the requirements he takes to be placed on him by his aims, this guidance will express itself in 'individual or joint practical deliberation about what to do, before or during the action, and in *ex post* justifications of his actions. In practical deliberation, the purpose provides the premise in the agent's deliberation, from which he proceeds to the conclusion that he should act in this way; and after the action it is to this aim that he appeals in justifying his action (as far as he is sincere)' (2011: 279).

According to Mayr, an agent's behaviour displays the structure characteristic of 'purposefulness' when facts of these three sorts obtain. Mayr claims that it is not necessary that facts of *all* three sorts obtain for an agent to act for a reason. Mayr thinks that sometimes an agent may not deliberate about what to do before acting, may be at a loss when asked later why he acted as he did,

have no success and failure feelings, and yet still act for a reason. For example, someone who has an unconscious (or subconscious) desire to sabotage a rival might give them bad advice. In this case, the agent has an aim (to sabotage his rival), but does not deliberate, would not be able to give an *ex post* justification for his action, and might not feel satisfied once the sabotage has been achieved. According to Mayr, 'what is present in such cases is only the (actual or hypothetical) teleological structure of the agent's behaviour' (2011: 282). Mayr thinks this indicates that facts of the first type are privileged in the sense that where an agent is acting with an aim, facts of the first type *must* obtain—something that doesn't hold true for the second or third type of facts.

9.2.2 Expanding on Mayr's theory

There are two issues with Mayr's account I would like to discuss. First, not all intentional activities display a pattern as sophisticated as the one Mayr describes. Some intentional actions are not done for reasons. For example, when I skip just for the fun of it, I have no aim I want to achieve by skipping. In such cases, because I have no aim I want to achieve, I have no aim to use in practical deliberation. Furthermore, because there's nothing I want to achieve by skipping, there are no success or failure feelings.[46] It is also unclear that I would engage in actions that are conducive to overcoming obstacles when I skip just for the fun of it. When I skip just for fun, it is more than likely that should some obstacle to skipping occur—e.g. my path becomes blocked or dangerously slippy—I would just stop skipping. I am doing it just for fun after all, not to achieve anything, so I have no motivation to continue skipping when doing so becomes difficult. Similarly, some animal behaviour seems to be intentional, in a minimal sense, even though it does not display anything as sophisticated as Mayr's 'plasticity'. For example, it seems to me that, when a cat grooms itself, the grooming is intentional, but it doesn't seem that, had the cat's environment presented an obstacle to grooming—e.g. had it started to rain—the cat would try to overcome this obstacle and continue grooming itself. In such circumstances, the cat is as likely to run off and hunt for mice as it is to go inside and continue grooming itself there. Many animal actions are, I think, intentional, but few have as sophisticated a teleological structure as Mayr describes.

Second, Mayr endorses the idea that rationalising explanations 'explain actions by making them intelligible' and not by positing an event-causal link between the agent's action and an appropriate mental event (2011: 269). What's

[46] If I go to skip and suddenly find myself unable, this will no doubt incur negative feelings, but they are not obviously 'failure feelings'—I am more likely to feel surprised and possibly concerned that a skill I thought I had has suddenly disappeared!

more, Mayr seems to endorse a context-placing or structural view of rationalising explanations:

> When we understand acting for a reason as following a standard of success … it must be the function of reasons-explanations to locate the action within the structure constituted by the agent's behaviour, emotional responses, thoughts, and practical reasoning which is constitutive for following the relevant standard of success. (2011: 292)

Mayr thus agrees with Julia Tanney that rationalising explanations explain by situating an agent's action within a wider pattern of activity the agent is engaging in, which thereby makes the action expected. Mayr also seems to agree with Megan Fritts that rationalising explanations are a form of structural explanation: rationalising explanations explain by connecting the agent's action with a more far-reaching description of the agent's activities in the same way in which structural explanations connect an explanandum with the form of the system of which the explanandum is part.

However, Mayr also thinks that rationalising explanations are a kind of disposition-citing explanation (2011: 295). He claims that, when a rationalising explanation is offered, a 'certain item of behaviour is explained *as the manifestation of one of the dispositions* connected with the welter of material and counterfactual conditionals which are responsible for the characteristic structure of intentional agency' (2011: 294, emphasis added). Mayr claims that the power manifested in intentional action is a 'complex power to act in certain ways in specific situations'; it is a power *of the agent* to structure her own activities (which are exercises of her abilities to act), a power that is 'superimposed on the pre-existing active powers of the agent' (2011: 295). So, on Mayr's view, rationalising explanations do two things: (a) they place the action explained within a specific structure and (b) they explain an action as the manifestation of a special sort of power to structure one's own activities, a power that is 'superimposed' on the pre-existing active powers of the agent. The second issue with Mayr's account I want to draw attention to concerns how rationalising explanations can perform both roles, and where this special power of an agent to structure her own activities comes from.

In response to the first issue, one might simply insist that actions like skipping for the fun of it and animal actions are not intentional because they do not meet the criteria Mayr sets out. However, even though actions like skipping for the fun of it and animal actions do not display a teleological structure as complex as the one Mayr describes, it is not true that they display no teleological structure at all. Anyone who can skip is able to make all sorts of small adjustments to their movements to maintain balance, or to ensure that the steps and hops that constitute skipping are executed with the required coordination. Skipping still involves some 'corrective measures', albeit on a smaller scale than the kind of corrective measures Mayr talks about. Similarly,

when a cat grooms itself, it must coordinate the movements of its body so that its tongue catches its fur in just the right way. Again, there is a form of teleological structure demonstrated. In both cases, there is a pattern demonstrated by the agent's actions—a pattern that makes sense once one learns what the agent is trying to do. I think that it is more in keeping with Mayr's core claim, that what makes an activity intentional is its characteristic structure of 'purposefulness', to grant that actions like skipping for the fun of it and animal actions are intentional in virtue of the teleological structure they display than to insist that such actions do not count as intentional because they fail to demonstrate a teleological structure of the right level of sophistication. If we are content to depart from traditional theories of intentional action and instead adopt a theory that ties the intentionality of some activity to the plasticity of that activity, then why not also accept the phenomenon of intentionality itself is not a homogenous phenomenon but instead something that can be more or less sophisticated?

The difficulty with weakening Mayr's view so that all activities that display some degree of plasticity count as intentional is that plasticity can be displayed in the behaviour of things that do not really act intentionally, for example machines and robots. This difficulty parallels issues surrounding Daniel Dennett's (1987) intentional stance theory. Dennett proposed that treating objects as rational agents with beliefs and desires helps us understand and predict the behaviour of those objects. Treating objects as rational agents with beliefs and desires is to take an intentional stance with respect to that object. According to Dennett, 'any object—or as I shall say, any *system*—whose behaviour is well predicted by this strategy is in the fullest sense of the word a believer' (1987: 15). Dennett goes on: '*What it is* to be a true believer is to be an *intentional system*, a system whose behaviour is reliably and voluminously predictable via the intentional strategy' (1987: 15, emphasis in original). The problem with Dennett's theory is that we can take the intentional stance to objects that do not really have beliefs and desires, like machines and robots.

It is commonly thought that there is a difference between *really* believing something and behaving *as if* you believed something, and that the difference lies in there being something extra, something hidden, in the case of genuine belief. I think this is the wrong way to capture the difference. True, machines and robots do not really have beliefs and desires, but this is not because believing something is a peculiar kind of property, or involves engaging in a peculiar kind of process. Rather, it is because machines and robots do not possess and exercise *two-way powers*. Their behaviour is not up to them. There is a real difference between behaviour of machines that seem to instantiate a pattern that can be made sense of by attributing mental states and genuine intentional action, but the difference does not consist in there being something *extra* present in the latter case. The difference is that machines are not capable of intentional action, because they do not possess two-way powers,

and possessing and exercising a two-way power is a necessary condition for acting intentionally.

A consideration that supports the idea that intentional agency always involves the exercise of two-way power is the fact that when an agent is constrained so that they only have the opportunity to φ, and lack the opportunity to not φ, if the agent φs in this situation we wouldn't want to say they φed intentionally.[47] For example, suppose Ben's hands have been temporarily paralysed so that he is denied the opportunity to move his hands. Whether Ben moves his hands or not is not up to him. Is it possible for Ben, in this situation, to intentionally refrain from moving his hands? Suppose someone unaware of Ben's situation said to him, "If you keep your hands perfectly still I'll give you £10." Ben may want to comply but, even if not moving is what Ben wants, it does not seem like he is remaining still intentionally when his hands are paralysed. It seems like being able to both move and not move your hands is a precondition for doing one or the other intentionally, and lacking this two-way power renders intentionally doing one or the other action impossible.

Another consideration that speaks in favour of the view that exercising a two-way power is a necessary condition for intentional action are cases of deviant causation. As discussed above, if possessing and exercising a two-way power is a necessary condition for acting intentionally, then we can explain why there is no intentional action in deviant causal chain cases; in such cases an agent's mental states rob the agent of the relevant two-way power.

The idea that possessing and exercising a two-way power is a necessary condition for acting intentionally suggests a possible answer to the second problem facing Mayr's account. It is because we have two-way powers that our activities can demonstrate patterns of the kind Mayr describes. When we have two-way powers, it is up to us whether we perform the activities these two-way powers are powers to do. In virtue of this, the pattern our actions display is also up to us. This is where, I think, the special power of an agent to structure her own activities, the power that Mayr says is 'superimposed' on the pre-existing active powers of the agent, comes from. Because we have *many* two-way powers, we also have an extra power to organise our actions in such a way so as to meet our aims. The power to act intentionally is thus an emergent power—a

[47] Frankfurt cases (Frankfurt 1969) are thought to demonstrate that this claim is false, that an agent can intentionally φ, and indeed be morally responsible for φing, even when they could not have done otherwise. However, I would argue that even in Frankfurt cases the agents in question do, in fact, have the ability and opportunity not to φ. The presence of neuroscientists with fancy machinery, who could take control over an agent's body just in case they start to look like they might not φ by themselves, may foreclose the physical possibility that a φing won't happen, but these facts are not relevant to what is an open agential possibility for the agent.

power that emerges from our possessing two-way powers to act. Having such a power does not mean we will always use it—many exercises of two-way powers are not intentional, for example absent-minded fiddling. The power may also come in degrees: creatures whose powers are mostly two-way will be able to organise their activities into a greater variety of patterns than creatures whose powers are mostly one-way. This allows us to articulate one way in which human action and animal action differ. Human beings possess more two-way powers than animals, which is to say that a greater proportion of human agential powers are two-way. This allows human beings to organise their activities into more complex patterns to meet a wider variety of aims.

This view has interesting consequences for the question of what causal information rationalising explanations provide. First, the view grants that rationalising explanations are a form of disposition-citing explanation. Intentional actions are manifestations of a special sort of power, namely a power to organise one's activities in accordance with a certain form (a power that depends on having two-way powers to act), and the function of rationalising explanations is to tell us which form the agent was disposed to structure her activities in accordance with. In this way, rationalising explanations tell us that the agent's activities are manifestations of a disposition to engage in activities that fall within a certain structure. For example, "Beth is buying flour because she wants to make bread" tells us that Beth's flour-buying is a manifestation of a disposition to engage in activities that are conducive to making bread, i.e. of her special power to organise her activities in accordance with a pattern that will be deemed successful by Beth if it ends with a loaf of bread.

Second, the view allows that rationalising explanations are also context-placing or structural. On the view proposed, intentional actions are manifestations of a special power of agents to organise their activities into a pattern of determinate form. As Mayr proposes, to act intentionally is for one's behaviour to display a particular kind of structure, i.e. 'the characteristic structure of taking something as one's "standard of success and failure", or "of correctness and incorrectness"' (2011: 271). The mental concepts cited in rationalising explanations make the structure of an agent's intentional activity intelligible. When we explain Beth's buying flour by attributing to her a desire to make bread, the function of this mental concept is to show how Beth's buying flour is part of a larger pattern of activities that display the structure typically associated with 'wanting to make bread'. When we learn that Beth is buying flour because she wants to make bread, we learn that Beth's activity sits within a pattern of activity that might include buying yeast, feeling disappointed if the bread tin is missing, consulting a cookbook etc.

Third, the view can explain why determining whether rationalising explanations provide information relevant to the manipulation or control of an effect, and hence whether rationalising explanations are causal, is difficult. As I mentioned in Chapter 8, when you learn that some agent's activity is a

manifestation of her desire or an output of her rational capabilities, you learn that you might be able to alter her activity by altering what she believes about the world, or by changing her desires, perhaps by changing her environment, but more usually by reasoning with her, talking to her, or persuading her. However, learning this information only makes it the case that you *might* be able to alter the agent's activity. The view of intentional action sketched in this section allows us to explain why this is: reasoning with an agent in an attempt to prevent them from φing (or to get them to φ) doesn't take away the agent's two-way power to φ. Because her power to φ is two-way, it is up to her whether she φs or not. Of course, we can always control someone else's φing by removing their two-way power to φ, for example by tying them down so that they no longer have the opportunity to φ. But learning about the reasons and motives behind an agent's activity is not relevant for our exercising *this* kind of control over the agent. If learning about the reasons and motives behind an agent's activity is relevant for the manipulation or control of their behaviour at all, then it is relevant for a kind of control that leaves the agent's two-way powers intact.

Determining whether rationalising explanations provide information relevant to the manipulation or control of an effect is difficult because it is unclear whether this latter sort of control is a form of *causal* control. Is convincing someone to behave in some way to exercise a *causal* power? Is it to cause something to happen? These questions matter if, as I have proposed, an explanation is causal if and only if it provides information relevant to manipulation and control, where manipulation and control are causal activities that powerful particulars, such as ourselves, can undertake. I do not think that the causal concept sits comfortably with concepts like *convincing, persuading* and *reasoning with*. On the other hand, the concept does not feel wholly inappropriate either. In short, because the disposition manifested when an agent acts intentionally is one which is dependent on their having and exercising *two-way* powers, learning about the reasons and motives behind an agent's activity does not provide us with information that enables us to *ensure* that the activity is (or is not) engaged in. However, it is not obvious that exercising causal control over a situation is always a matter of *ensuring* certain outcomes. The causal status of rationalising explanations is atypical. But if something like the account of intentional action I have sketched in this section is true, then the unique causal nature of rationalising explanations is not an anomaly; it is instead something that should be expected given the nature of the agential powers demonstrated in intentional action.

In this chapter, I have proposed an alternative view of intentional actions, inspired by Mayr (2011), which takes intentional actions to be manifestations of a special power of agents to organise their activities into a pattern of determinate form (an emergent power that depends on one possessing two-way powers to act). Rationalising explanations reveal this form by attributing mental states with certain contents to the agent.

References[48]

Alvarez, M 2013 Agency and two-way powers. *Proceedings of the Aristotelian Society*, 113(1pt1): 101–121. DOI: https://doi.org/10.1111/pash.2013.113.issue-1pt1

Berofsky, B 2011 Compatibilism without Frankfurt: Dispositional analyses of free will. In: Kane, R *The Oxford handbook of free will*. 2nd ed. Oxford: Oxford University Press. pp. 153–174.

Bishop, J 1989 *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Campbell, J K 2005 Compatibilist alternatives. *Canadian Journal of Philosophy*, 35(3): 387–406. DOI: https://doi.org/10.1080/00455091.2005.10716595

Davidson, D 1963 Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700. DOI: https://doi.org/10.2307/2023177. Reprinted in Davidson 2001 pp. 3–20.

Davidson, D 1971 Agency. In: Binkley, R, Bronaugh, R and Marras, A *Agent, action, and reason*. Toronto: University of Toronto Press. pp. 1–37. Reprinted in Davidson 2001 pp. 43–62.

Davidson, D 1973 Freedom to act. In: Honderich, T *Essays on freedom of action*. New York: Routledge and Kegan Paul. pp. 137–156. Reprinted in Davidson 2001 pp. 63–82.

Davidson, D 1987 Problems in the explanation of action. In Smart, J J C et al. *Metaphysics and morality: Essays in honour of J.J.C. Smart*. New York: Blackwell. pp. 35–49.

Davidson, D 2001 *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Dennett, D 1987 *The intentional stance*. Cambridge, MA: MIT Press.

Frankfurt, H 1969 Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23): 829–839. DOI: https://doi.org/10.2307/2023833

Frost, K 2013 Action as the exercise of a two-way power. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 611–624. DOI: https://doi.org/10.1080/0020174x.2013.841043

Hornsby, J 2015 Causality and 'the mental'. *HUMANA.MENTE Journal of Philosophical Studies*, 8(29): 125–140.

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Kapitan, T 2011 A compatibilist reply to the consequence argument. In: Kane, R *The Oxford Handbook of Free Will*. 2nd ed. Oxford: Oxford University Press. pp. 131–150.

Kenny, A 1975 *Will, freedom and power*. Oxford: Basil Blackwell.

[48] Author note: some references to Davidson are formatted (1963/2001). This indicates the initial date of publication of the paper (in this case 1963) but references the paper as it appears in the 2001 collection of his essays, with the page numbers relating to that volume.

List, C 2014 Free will, determinism, and the possibility of doing otherwise, *Noûs*, 48(1): 156–178. DOI: https://doi.org/10.1111/nous.12019

Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.

Mele, A 2003 *Motivation and agency*. Oxford: Oxford University Press.

Smith, M 2012 Four objections to the standard story of action (and four replies). *Philosophical Issues*, 22(1): 387–401. DOI: https://doi.org/10.1111/j.1533-6077.2012.00236.x

Steward, H 2012 *A metaphysics for freedom*. Oxford: Oxford University Press.

Steward, H 2013a Responses. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 681-706. DOI: https://doi.org/10.1080/0020174x.2013.841055

Woodfield, A 1976 *Teleology*. Cambridge: Cambridge University Press.

CHAPTER 10

Mental Causation Reconsidered

In most discussions of the problem of mental causation, mental causation is presented as a cause–effect relation between mental and physical items. Mentality and physicality are presented as two sides of a causal exchange. I called this understanding of mental causation the relational understanding of mental causation.

> **Relational understanding of mental causation:** mental causation is mental items (events, processes or states) standing in causal relations to physical items (e.g. movements of a person's body).

Philosophers writing about the problem of mental causation are limited to this way of describing what mental causation is, because they assume that 'cause' is an unequivocal term—all causation everywhere is the same kind of thing, so the only thing that can discriminate between different categories of causation is the nature of the relata involved. What is 'mental' about mental causation is that it involves at least one mental relatum. I argued that this understanding of mental causation is a crucial component of the main argument for adopting a physicalist metaphysics of mind. However, it is my view that this is a flawed approach to understanding mental causation.

One of the aims of this book was to explain why the relational understanding of mental causation is presupposed in many debates in philosophy of mind. In the first three chapters, I showed that the relational understanding of mental causation is entailed by a triad of philosophical theories: physicalism, causal theories of intentional action and a relational approach to causation. I argued that, even though these theories are logically independent, in practice they reinforce each other. I called this triad the physicalist triad because the upshot of endorsing these three theories is that physicalism ends up seeming like the only possible metaphysics of mind that stands a chance of saving the phenomenon of mental causation.

How to cite this book chapter:
White, Andrea. 2024. *Understanding Mental Causation*. Pp. 197–204. York: White Rose University Press. DOI: https://doi.org/10.22599/White.k.

My second aim in this book was to try to describe a way to break out of the physicalist triad. In so doing, I hoped to break physicalism's hegemony over our thinking about the mind. The strategy I followed was to focus on what I take to be the weakest element of the physicalist triad, namely its account of human agency. The physicalist triad entails a physicalist/event-causalist description of human agency, where what it is to act is to do something intentionally, and what it is for an action to be intentional is explained in terms of causation by a mental state of the agent, or a mental event involving the agent. And, according to physicalism, these mental items are realised by physical items—most plausibly neural events, or perhaps physical events that are themselves complex and include neural events as parts. The picture of human agency that emerges is a reductive one. What it is for a person to act is nothing more than the triggering of bodily movements by sub-personal events. This picture of human agency is endorsed, at least partially, by Bishop (1989), Brand (1984), Bratman (1987), Dretske (1988), Enç (2003), Mele (1992; 2003) and Shepherd (2021).

The problem with this physicalist/event-causal picture of agency is that, when causal reality is viewed as nothing but chains of causally related events, everything in the causal world is something that occurs or something that happens. Occurrences and happenings are not things that anyone 'does'. So, when causal reality is viewed as nothing but chains of causally related events, the agent does not seem like an agent anymore, because the agent does not seem to do anything; they seem instead to be merely the setting for events to cause other events. This is the disappearing agent objection, which essentially says that there is something about our concept of agency and something about the idea of the causal world as consisting of nothing but chains of causally related events that don't marry: agency is about agents doing things; a causally related chain of events contains only what occurs or happens. The disappearing agent objection is often dismissed as either begging the question against the physicalist/event-causal account of agency or merely showing that standard physicalist/event-causal accounts needs to be modified to include a causal sequence that plausibly plays the functional role of the agent, or only being a problem for libertarian accounts of free will. However, I believe the disappearing agent objection should be taken seriously: there really is a kind of incompatibility between our concept of agency and the idea of the causal world as consisting of nothing but chains of causally related events.

The disappearing agent objection should be taken seriously because the boundary between agential and non-agential does not map onto the divide between event-causal sequences that involve intentional states and those that do not. Sometimes a certain kind of causation by a mental state is what stops an event counting as an instance of agency (deviant causal chain cases); our agency concept extends to cases where agents remain passive and so there is no action to be caused; and our concept of agency extends to cases where there is no mental cause of a bodily movement. What this suggests is that attempting to understand agency in terms of a distinction between event-causal

sequences that involve intentional states and those that do not misconstrues the agency concept.

I concluded that, to properly understand agency, what is needed is a radical departure from the physicalist triad, and in particular the relational approach to causation. Specifically, to understand agency, we need a metaphysical framework that allows us to think of causation as something other than a relation between events. Only then is it possible to see how the causality of action might be something other than a causal relation between mental event and action, and instead something that casts the agent as a causal player, rather than merely the setting for events to cause other events.

In Chapter 6, I outlined a non-relational approach to causation. According to this approach, causation is not always and everywhere a relation, and giving a full account of causation is not merely a matter of explaining what a relation must be like to be a causal relation. Put positively, I maintain that causation can be a process rather than a relation, of which processes like breaking, crushing, bending etc. are more determinate species. My process ontology maintains that processes are universals that substances engage in, and events are instances of processes—they are particular occurrences that come into being when a substance has engaged in a process and completed it.

I argued in Chapter 9 that this non-relational approach to causation, and the process ontology that accompanies it, allows us to put together a more successful understanding of agency. On my view, agents are substances that exercise agential powers, where to exercise a power is for a substance to engage in a process, i.e. for a dynamic state of affairs to obtain. On this view, like other agent-causal accounts of agency, agency is a kind of causation where the agent, who is taken to be a substance, exercises causal power and this exercise of causal power cannot be reduced to causation by an event involving the agent. What makes an action a demonstration of agency is that *the agent* is causing something to happen, where this *causing* of the agent cannot be understood as the causation of one event by another—it is its own special type of causation. However, unlike other agent-causal accounts, I propose that the special type of causation demonstrated in agency is a *process*—not a relation. What it is for a substance to be causing something is for there to be an *activity*—i.e. a way for substances to be effecting change—which the substance is engaging in. Actions are the events that come into existence when agents exercise their agential powers—i.e. engage in processes—and then complete those processes.

I also argued that there are two distinctions crucial to our concept of agential power: the distinction between activity and passivity, and the distinction between one-way and two-way powers. Agency does not reduce to the exercise of active power, because some substances can manifest their agency by remaining passive, and therefore by not engaging in activity. Neither does agency reduce to the exercise of two-way power, because not all substances that cause things to happen do so by exercising two-way powers, but all substances that cause things to happen are agents. My view is that agency is a complex concept

that incorporates both distinctions. Some substances' agential powers are one-way; these substances manifest their agency when they are active but not when they are passive. Other substances' agential powers are two-way; these substances manifest their agency when they are active, but also sometimes when they are passive.

My non-relational approach to causation also opened up new ways of understanding intentional action. Many philosophers have tried to provide an account of intentional action by examining the distinctive sort of explanation with which intentional actions are associated, i.e. rationalising explanations. Davidson (1963) argues that rationalising explanations are causal explanations. They are true if a mental event suitably related to the mental concept cited in the rationalising explanation stands in a causal relation to the action explained. Davidson's argument that rationalising explanations are causal is often taken to justify the claim that mental states or events stand in causal relations to intentional actions. Thus, Davidson's argument is the source of the common view that our conception of ourselves as intentional agents presupposes that mentality is causally relevant in the physical world and that this mental causation should be conceived of in relational terms.

In Chapters 7 and 8, I challenged Davidson's argument that states of desiring and states of believing are causes of the actions they explain. I argued that it is not necessary for an explanation to be causal that its explanandum designate an effect and its explanans designate an item that is the cause of that effect. My non-relational theory of causation implies that facts about causal relations between events are not the only causal facts that causal explanations could answer to. Some causal explanations are made true by the non-relational aspect of causal reality, that is, by facts about substances engaging in processes.

Explanations of intentional action that cite the agent's reasons for acting are the kind of causal explanation that is not made true by causally related events. The most important consideration favouring this view is that it saves two strong intuitions: (a) that reason-giving explanations are causal, and (b) that the mental states cited in reason-giving explanations do not denote items that stand in causal relations to the actions they explain. The second intuition is bolstered by the many arguments offered by non-causalists, which are discussed in Chapter 7, that rationalising explanations need not be considered causal in Davidson's sense to meet Davidson's challenge. The idea that rationalising explanations are causal explanations that answer to the non-relational aspect of causal reality is also supported by the fact that rationalising explanations bear some similarities to both process-citing and disposition-citing explanations.

If these arguments are successful, they show that the fact that we causally explain people's intentional actions by referencing (sometimes directly, sometimes indirectly) their mental states does not justify the contention that, necessarily, whenever there is intentional action there is a causal relation between a mental item and an action or bodily movement. When we say that someone acted intentionally because of what she believed, desired, intended or decided,

these mental concepts need not refer to items that stand in causal relations to physical events. The causal nature of rationalising explanations does not give us any reason to think that there are causal relations between mental items and physical items whenever we act intentionally.

This view, that rationalising explanations are causal explanations that do not designate mental items that stand to the action explained as cause to effect, has consequences for how we ought to think about the nature of intentional action. Most importantly, it casts doubt on the view that intentional actions are distinguished from non-intentional actions by their causes. In Chapter 9, I proposed an alternative view of intentional actions, inspired by Mayr (2011). I proposed that to act intentionally is to engage in a process, and as such is to exercise a power—but a power of a special sort. Intentional actions are manifestations of a special power to organise one's activities into a pattern of determinate form. This power emerges from our possessing two-way powers to act: because we have *many* two-way powers, we also have an extra power to organise our actions into patterns. Rationalising explanations reveal the form of this pattern by attributing mental states with certain contents to the agent. In this way, rationalising explanations are context-placing or structural because they reveal the structure of our activities and make our activities intelligible by helping us see that they are part of a larger pattern of activity. However, rationalising explanations are also disposition-citing because the function of rationalising explanations is to tell us which form the agent was disposed to structure her activities in accordance with.

Is there anything worthy of the name 'mental causation' necessarily on display whenever an agent acts intentionally? I believe we can, and should, answer this question positively. I have mentioned that it is natural to think that some form of mental causation, or 'the reality of causal processes involving cognitive phenomena' as Peter Menzies (2013: 58) puts it, is indispensable to our conception of ourselves as agents who act intentionally and bear moral responsibility. A positive answer to this question is possible once we acknowledge that we need not, and should not, understand 'mental' in 'mental causation' as a 'transferred epithet', as Tim Crane (1995: 219) puts it. Understanding 'mental' in 'mental causation' as qualifying the cause relatum of a causal relation, rather than causation itself, is a prescription of the relational understanding of mental causation.

An alternative conception of the mentality of the causal processes human beings engage in when they act intentionally is that it consists in the fact that these processes are part of a larger pattern of *meaningful*, or *interpretable*, activity.

I have proposed that acting intentionally is to manifest a special power to organise one's activities into a pattern that can be made sense of by appeal to mental concepts. When an agent acts intentionally, the activity the agent is engaging in is part of a larger teleological structure whose form is revealed by attributing knowledge, beliefs, desires or aims to the agent. Furthermore, when you learn that some agent's activity is a manifestation of her desire or an output

of her rational capabilities, you learn that you might be able to alter her activity by altering what she believes about the world, or by changing her desires, usually by reasoning with her, talking to her or persuading her. However, learning this information only makes it the case that you *might* be able to alter the agent's activity. This is because reasoning with an agent in an attempt to prevent them from φing (or get them to φ) doesn't take away the agent's two-way power to φ, so it remains up to her whether she φs or not. Learning about an agent's reasons for acting therefore allows one to manipulate and control the agent's behaviour in a unique way: in a way that leaves the agent's two-way powers intact. I suggest that these are the facts about intentional action that make the causation an agent engages in when they act intentionally count as 'mental'.

Acting intentionally is not mental causation because it consists in actions caused to happen by mental events. Acting intentionally is mental causation in virtue of the fact that the causal activities agents engage in when they act intentionally are part of a larger teleological structure whose form is revealed by attributing mental states to the agent and which we can manipulate in a unique way, i.e. using reasoning and persuasion. For example, the mental causation that is on display when I add salt to the sauce because I think it will make it taste better does not consist in causation of my hand movements by some mental item—e.g. a belief that adding salt will make the sauce taste better. Instead, the causal processes I engage in count as mental causation in virtue of the fact that this particular activity (adding salt to the sauce) is part of a larger pattern of activity whose form and typical trajectory is revealed when it is understood that I want to make the sauce taste better and believe adding salt will achieve that. It is also mental causation in virtue of the fact that persuading me that something else would improve the sauce more effectively is a means by which someone could alter the trajectory of my behaviour while leaving my two-way powers to act intact.

Does my suggestion really capture our intuitive understanding of what mental causation is? Thomas Kroedel (2020) suggests that mental causation can be summarised as the idea that what's going on in your mind makes a difference to what's going on in the world, which is to say that, had our minds been different, our activities would be too. What seems undeniable is that our mental life makes a difference to our bodily life: what we think, what we believe, what we want, what we feel affects what we do with our bodies. It has been a mistake, I think, to interpret this pre-philosophical view as claiming that there is causal interaction between mind and body. The Cartesian notion that our mental life affects what we do with our bodies because we have a mind that causes our body to move is incorrect.

I also think that understanding mental causation as a causal exchange between distinct aspects of ourselves (the mental and the physical) is incorrect. The influence of our mentality on our activities does not reduce to events inside us triggering bodily movements. However, I do not think this is the only way to interpret the naïve idea that what's going on in your mind makes a difference

to what we do with our bodies. Instead, my suggestion is that our minds make a difference to what is going on in the world because we make a difference. When we act intentionally, that is our minds making a difference to the world. Our mental life makes a difference to our bodily life because we have the power to organise our activities into patterns that are made comprehensible by our mental states.

Debates within philosophy of mind tend to centre on which metaphysics of mind best reconciles the claim that mental items stand in causal relations to physical events with plausible principles about what actual causation is like, such as the principle of causal closure. However, if realism about mental causation does not require the relational understanding of mental causation at all, then the problem of mental causation as it is standardly understood may be a pseudo-problem.

Human beings are capable of performing activities that we would naturally describe as 'mental', such as imagining and reasoning, and persuading and convincing. Exactly what these activities amount to is a difficult philosophical question. However, it seems to me that these activities are ways to deliberate—individually or in groups—about what beliefs and desires it is best to have, and can be means by which we alter what beliefs or desires an agent has. That we have such capacities is relevant to our bearing moral responsibility.

How it is that we have such capacities is, I think, a very difficult question. How are we able to engage in activities like imagining and reasoning? How does our capacity to imagine, reason, persuade or convince relate to the physical capacities of our bodies? How is it possible that we can change the action plans and projects an agent is disposed to enact by imagining or reasoning or persuading or convincing? I have no idea how to answer these questions. But it is *these* questions—and not questions about how mental items can stand in causal relations to physical events—that constitute the real problem of mental causation. The real mystery is not how mental items can stand in causal relations to physical events but how it is that we can perform mental activities at all.

References

Bishop, J 1989 *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Brand, M 1984 Intending and acting. *Mind*, 96(381): 121–124.

Bratman, M 1987 *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Crane, T 1995 The mental causation debate. *Aristotelian Society Supplementary Volume*, 69 (1): 211–236. DOI: https://doi.org/10.1093/aristoteliansupp/69.1.211

Dretske, F 1988 *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Enç, B 2003 *How we act: Causes, reasons, and intentions*. New York: Oxford University Press.
Kroedel, T 2020 *Mental causation: A counterfactual theory*. Cambridge: Cambridge University Press.
Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.
Mele, A 1992 *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.
Mele, A 2003 *Motivation and agency*. Oxford: Oxford University Press.
Menzies, P 2013 Mental causation in the physical world. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 58–88.
Shepherd, J 2021 *The shape of agency: Control, action, skill, knowledge*. Oxford: Oxford University Press.

Works Cited

Alvarez, M 2010 *Kinds of reasons: An essay in the philosophy of action*. Oxford: Oxford University Press.

Alvarez, M 2013 Agency and two-way powers. *Proceedings of the Aristotelian Society*, 113(1pt1): 101–121. DOI: https://doi.org/10.1111/pash.2013.113.issue-1pt1

Alvarez, M and Hyman, J 1998 Agents and their actions. *Philosophy*, 73(284): 219–245. DOI: https://doi.org/10.1017/s0031819198000199

American Honda Motor Co. Ltd. Public Relations Division 2007 ASIMO Technical Manual, September 2007. Available at https://asimo.honda.com/downloads/pdf/asimo-technical-information.pdf [Last accessed 3 September 2023].

Anscombe, G E M 1957 *Intention*. Oxford: Basil Blackwell.

Anscombe, G E M 1971 *Causality and determination: An inaugural lecture*. Cambridge: Cambridge University Press.

Anscombe, G E M 2000 *Intention*. Cambridge, MA: Harvard University Press. Originally published in England in 1957 by Basil Blackwell.

Armstrong, D M 1968 *A materialist theory of the mind*. London: Routledge.

Armstrong, D M 1978a *Universals and scientific realism*. New York: Cambridge University Press.

Armstrong, D M 1978b *A theory of universals. Universals and scientific realism volume II*. New York: Cambridge University Press.

Armstrong, D M 1989 *Universals: An opinionated introduction*. Boulder, CO: Westview Press.

Armstrong, D M 1997 *A world of states of affairs*. Cambridge: Cambridge University Press.

Árnadóttir, S T and Crane, T 2013 There is no exclusion problem. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 248–266.

Bach, K 1980 Actions are not events. *Mind*, 89(353): 114–120. DOI: https://doi.org/10.1093/mind/lxxxix.353.114

Baumgartner, M 2008 Regularity theories reassessed. *Philosophia*, 36(3): 327–354. DOI: https://doi.org/10.1007/s11406-007-9114-4

Beebee, H 2004 Causing and nothingness. In: Collins, J, Hall E J and Paul, L A, *Causation and counterfactuals*. Cambridge, MA: MIT Press. pp. 291–308.

Beebee, H 2006 Does anything hold the universe together? *Synthese*, 149(3): 509–533. DOI: https://doi.org/10.1007/s11229-005-0576-2

Beebee, H 2007 The two definitions and the doctrine of necessity. *Proceedings of the Aristotelian Society*, 107(1pt3): 413–431. DOI: https://doi.org/10.1111/j.1467-9264.2007.00231.x

Bennett, J 1988 *Events and their names*. Indianapolis, IN: Hackett.

Bennett, K 2003 Why the exclusion problem seems intractable and how, just maybe, to tract it. *Noûs*, 37(3): 471–497. DOI: https://doi.org/10.1111/1468-0068.00447

Berofsky, B 2011 Compatibilism without Frankfurt: Dispositional analyses of free will. In: Kane, R *The Oxford handbook of free will*. 2nd ed. Oxford: Oxford University Press. pp. 153–174.

Bishop, J 1989 *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Bokulich, A 2011 How scientific models can explain. *Synthese*, 180(1): 33–45. DOI: https://doi.org/10.1007/s11229-009-9565-1

Brand, M 1984 Intending and acting. *Mind*, 96(381): 121–124.

Bratman, M 1987 *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Bratman, M 2001 Two problems about human agency. *Proceedings of the Aristotelian Society*, 101(3): 309–326. DOI: https://doi.org/10.1111/j.0066-7372.2003.00033.x

Brent, M 2017 Agent causation as a solution to the problem of action. *Canadian Journal of Philosophy*, 47(5): 656–673. DOI: https://doi.org/10.1080/00455091.2017.1285643

Bryant, A 2020 Physicalism. In: Raven, M J *The Routledge handbook of metaphysical grounding*. New York: Routledge. pp. 484–500.

Burge, T 1979 Individualism and the mental. *Midwest Studies in Philosophy*, 4(1): 73–122. DOI: https://doi.org/10.1111/j.1475-4975.1979.tb00374.x

Campbell, J K 2005 Compatibilist alternatives. *Canadian Journal of Philosophy*, 35(3): 387–406. DOI: https://doi.org/10.1080/00455091.2005.10716595

Cartwright, N 2009 Causal laws, policy predictions, and the need for genuine powers. In: Handfield, T *Dispositions and causes*. Oxford: Oxford University Press.

Chalmers, D 1996 *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Chalmers, D 2020 Is the hard problem of consciousness universal? *Journal of Consciousness Studies*, 27(5–6): 227–257.

Child, W 1994 *Causality, interpretation, and the mind*. New York: Oxford University Press.

Chisholm, R 1964 Human freedom and the self. In Kane, R *Free will*. Malden, MA: Wiley Blackwell.

Chisholm, R 1976 *Person and object: A metaphysical study*. London: Routledge.

Clarke, R 2003 *Libertarian accounts of free will*. New York: Oxford University Press.

Clarke, R 2014 *Omissions: Agency, metaphysics, and responsibility*. New York: Oxford University Press.

Clarke, R 2017 Free will, agent causation, and 'disappearing agents'. *Noûs*, 53(1): 76–96. DOI: https://doi.org/10.1111/nous.12206

Coope, U 2007 Aristotle on action. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 81: 109–138. DOI: https://doi.org/10.1111/j.1467-8349.2007.00153.x

Crane, T 1995 The mental causation debate. *Aristotelian Society Supplementary Volume*, 69(1): 211–236. DOI: https://doi.org/10.1093/aristoteliansupp/69.1.211

Crane, T and Mellor, D H 1990 There is no question of physicalism. *Mind*, 99(394): 185–206. DOI: https://doi.org/10.1093/mind/xcix.394.185

Crowther, T 2011 The matter of events. *Review of Metaphysics*, 65(1): 3–39.

Curry, D 2018 Beliefs as inner causes: The (lack of) evidence. *Philosophical Psychology*, 31(6): 850–877. DOI: https://doi.org/10.1080/09515089.2018.1452197

D'Oro, G 2012 Reasons and causes: The philosophical battle and the metaphilosophical war. *Australasian Journal of Philosophy*, 90(2): 207–221. DOI: https://doi.org/10.1080/00048402.2011.583930

Dancy, J 2000 *Practical reality*. New York: Oxford University Press.

Davidson, D 1963 Actions, reasons, and causes. *Journal of Philosophy*, 60(23): 685–700. DOI: https://doi.org/10.2307/2023177. Reprinted in Davidson 2001a pp. 3–20.

Davidson, D 1967 Causal relations. *Journal of Philosophy*, 64(21): 691–703. Reprinted in Davidson 2001a pp. 149–162.

Davidson, D 1970 Mental events. In: Foster, L and Swanson, J W *Experience and theory*. 2nd ed. Cambridge, MA: University of Massachusetts Press. Reprinted in Davidson 2001a pp. 207–224.

Davidson, D 1971 Agency. In: Binkley, R, Bronaugh, R and Marras, A *Agent, action, and reason*. Toronto: University of Toronto Press. pp. 1–37. Reprinted in Davidson 2001a pp. 43–62.

Davidson, D 1973 Freedom to act. In: Honderich, T *Essays on freedom of action*. New York: Routledge and Kegan Paul. pp. 137–156. Reprinted in Davidson 2001a pp. 63–82.

Davidson, D 1982 Rational animals. *Dialectica*, 36(4): 317–328. Reprinted in Davidson 2001b pp. 95–106.

Davidson, D 1987 Problems in the explanation of action. In Smart, J J C et al. *Metaphysics and morality: Essays in honour of J.J.C. Smart*. New York: Blackwell. pp. 35–49.

Davidson, D 1997 Indeterminism and antirealism, In: Kulp, C B *Realism/antirealism and epistemology*. Lanham, MD: Rowman & Littlefield. pp. 109–122. Reprinted in Davidson 2001b pp. 69–84.

Davidson, D 2001a *Essays on actions and events*. 2nd ed. Oxford: Clarendon Press.

Davidson, D 2001b *Subjective, intersubjective, objective*. Oxford: Clarendon Press.

De Swart, H 1996 Quantification over time. In: van der Does, J and van Eijck, J *Quantifiers, logic, and language*. Cambridge: Cambridge University Press.

De Swart, H 2012 Verbal aspect. In: Binnick, R I *The Oxford handbook of tense and aspect*. New York: Oxford University Press. pp. 752–781.

Dennett, D 1987 *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D 2005 *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.

DiYanni, C and Kelemen, D 2005 Using a bad tool with good intention: How preschoolers weigh physical and intentional cues when learning about artifacts. *Cognition*, 97: 327–335. DOI: https://doi.org/10.1016/j.jecp.2008.05.002

Dretske, F 1988 *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Ehring, D 2011 *Tropes: Properties, objects, and mental causation*. Oxford: Oxford University Press.

Elpidorou, A and Dove, G 2018 *Consciousness and physicalism: A defense of a research program*. New York: Routledge.

Enç, B 2003 *How we act: Causes, reasons, and intentions*. New York: Oxford University Press.

Fales, E 1990 *Causation and universals*. London: Routledge.

Fifel, K 2018 Readiness potential and neuronal determinism: New insights on Libet experiment. *Journal of Neuroscience*, 38(4): 784–786. DOI: https://doi.org/10.1523/JNEUROSCI.3136-17.2017

Filip, H 2012 Lexical aspect. In: Binnick, R I *The Oxford handbook to tense and aspect*. New York: Oxford University Press. pp. 721–752.

Flanagan, O J 1996 *Self expressions: Mind, morals, and the meaning of life*. Oxford: Oxford University Press.

Fodor, J 1987 *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Frankfurt, H 1969 Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23): 829–839. DOI: https://doi.org/10.2307/2023833

Frankfurt, H 1978 The problem of action. *American Philosophical Quarterly*, 15(2): 157–162.

Fritts, M 2021 Reasons explanations (of actions) as structural explanations. *Synthese*, 199(5–6): 12683–12704. DOI: https://doi.org/10.1007/s11229-021-03349-4

Frost, K 2013 Action as the exercise of a two-way power. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 611–624. DOI: https://doi.org/10.1080/0020174x.2013.841043

Gallow, J 2022 The metaphysics of causation, The Stanford Encyclopedia of Philosophy, 14 April 2022. Available at https://plato.stanford.edu/archives/sum2022/entries/causation-metaphysics [Last accessed 2 September 2023].

Galton, A 2018 Processes as patterns of occurrence. In: Stout, R *Process, action, and experience*. Oxford: Oxford University Press.

Ganeri, J, Noordhof, P and Ramachandran, M 1996 Counterfactuals and preemptive causation. *Analysis*, 56(4): 219–225. DOI: https://doi.org/10.1093/analys/56.4.219

Gibb, S 2013 Introduction to mental causation and ontology. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 1–17.

Gvozdanović, J 2012 Perfect and imperfect aspect. In: Binnick, R I *The Oxford handbook to tense and aspect*. New York: Oxford University Press. pp. 781–803.

Hacker, P 2007 *Human nature*. Oxford: Blackwell.

Haddock, A 2005 At one with our actions, but at two with our bodies: Hornsby's account of action. *Philosophical Explorations*, 8(2): 157–172. DOI: https://doi.org/10.1080/13869790500095939

Haggard, P and Libet, B 2001 Conscious intention and brain activity. *Journal of Consciousness Studies*, 8(11): 47–63.

Harré, R and Madden, E H 1975 *Causal powers*. Oxford: Blackwell.

Hart, H L and Honoré, A M 1985 *Causation in the law*. 2nd ed. Oxford: Oxford University Press.

Haslanger, S 2016 What is a (social) structural explanation? *Philosophical Studies*, 173(1): 113–130. DOI: https://doi.org/10.1007/s11098-014-0434-5

Hatano, G and Inagaki, K 1994 Young children's naive theory of biology. *Cognition*, 50(1–3): 171–188. DOI: https://doi.org/10.1016/0010-0277(94)90027-2

Haugeland, J 1982 Weak supervenience. *American Philosophical Quarterly*, 19(1): 93–103. http://www.jstor.org/stable/20013945

Haynes, J-D and Pauen, M 2013 The complex network of intentions. In Caruso, G D *Exploring the illusion of free will and moral responsibility*. Plymouth: Lexington Books. pp. 221–238.

Heil, J 2013 Mental causation. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 18–35.

Hellman, G P and Thompson, F W 1975 Physicalism: Ontology, determination, and reduction. *Journal of Philosophy*, 72(17): 551–564. DOI: https://doi.org/10.2307/2025067

Hempel, C 1969 Reduction: Ontological and linguistic facets. In White, M et al. *Philosophy, science, and method: Essays in honor of Ernest Nagel*. New York: St Martin's Press. pp. 179–199.

Hempel, C and Oppenheim, P 1948 Studies in the logic of explanation. *Philosophy of Science*, 15(2): 135–175. DOI: https://doi.org/10.1086/286983

Higginbotham, J 2000 On events in linguistic semantics. In: Higginbotham, J, Pianesi F and Varzi, A *Speaking of events*. New York: Oxford University Press. pp. 49–81.

Hill, C S 1997 Imaginability, conceivability, possibility and the mind-body problem. *Philosophical Studies*, 87(1): 61–85. DOI: https://doi.org/10.1023/a:1017911200883

Hinshelwood, A 2013 The metaphysics and epistemology of settling: Some Anscombean reservations. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 625–638. DOI: https://doi.org/10.1080/0020174x.2013.841044

Hocutt, M 1974 Aristotle's four becauses. *Philosophy*, 49(190): 385–399. DOI: https://doi.org/10.1017/s0031819100063324

Holland, P W 1986 Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960. DOI: https://doi.org/10.2307/2289064

Hopkins, J 1978 Mental states, natural kinds and psychophysical laws. *Aristotelian Society Supplementary Volume*, 52(1): 195–236. DOI: https://doi.org/10.1093/aristoteliansupp/52.1.195

Horgan, T E 1993 From supervenience to superdupervenience: Meeting the demands of a material world. *Mind*, 102(408): 555–586. DOI: https://doi.org/10.1093/mind/102.408.555

Hornsby, J 1980 *Actions*. London: Routledge.

Hornsby, J 2004a Agency and alienation. In: Macarthur, D and De Caro, M *Naturalism in question*. Cambridge, MA: Harvard University Press. pp. 173–187.

Hornsby, J 2004b Agency and actions. In: Steward, H and Hyman, J *Agency and action*. Cambridge: Cambridge University Press. pp. 1–23.

Hornsby, J 2012 Actions and activity. *Philosophical Issues*, 22(1): 233–245. DOI: https://doi.org/10.1111/j.1533-6077.2012.00227.x

Hornsby, J 2015 Causality and 'the mental'. *HUMANA.MENTE Journal of Philosophical Studies*, 8(29): 125–140.

Hume, D 1964 *A treatise of human nature. Volume 1*. London: J M Dent/E P Dutten.

Hume, D 1975 *Enquiries concerning human understanding and concerning the principles of morals*. 3rd ed. Oxford: Clarendon Press.

Hursthouse, R 1991 Arational actions. *Journal of Philosophy*, 88(2): 57–68. DOI: https://doi.org/10.2307/2026906

Hyman, J 2015 *Action, knowledge, and will*. New York: Oxford University Press.

Jackson, F 1995 Essentialism, mental properties and causation. *Proceedings of the Aristotelian Society*, 95: 253–268. DOI: https://doi.org/10.1093/aristotelian/95.1.253

Jackson, F 1998 *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Kapitan, T 2011 A compatibilist reply to the consequence argument. In: Kane, R *The Oxford Handbook of Free Will*. 2nd ed. Oxford: Oxford University Press. pp. 131–150.

Kenny, A 1975 *Will, freedom and power*. Oxford: Basil Blackwell.

Kim, J 1976 Events as property exemplifications. In: Brand, M and Walton, D *Action theory*. Dordrecht: D. Reidel. pp. 310–326.

Kim, J 1989 The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63(3): 31–47. DOI: https://doi.org/10.2307/3130081

Kim, J 1993 *Supervenience and the mind: Selected essays*. Cambridge: Cambridge University Press.

Kim, J 1998 *Mind in a physical world*. Cambridge: Cambridge University Press.

Kim, J 2001 Mental causation and consciousness: The two mind-body problems for the physicalist. In: Carl, G and Loewer, B *Physicalism and its discontents*. Cambridge: Cambridge University Press. pp. 271–283.

Kim, J 2005 *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.

Kitcher, P 1989 Explanatory unification and the causal structure of the world. In Kitcher, P and Salmon, W *Scientific explanation*. Minneapolis, MN: University of Minnesota Press. pp. 410–505.

Kroedel, T 2020 *Mental causation: A counterfactual theory*. Cambridge: Cambridge University Press.

Levine, J 1983 Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(October): 354–361. DOI: https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Levy, N 2005 Libet's impossible demand. *Journal of Consciousness Studies*, 12(12): 67–76.

Lewis, D K 1973a Causation. *The Journal of Philosophy*, 70(17): 556–567. DOI: https://doi.org/10.2307/2025310

Lewis, D K 1973b *Counterfactuals*. Oxford: Basil Blackwell.

Lewis, D K 1986 *Philosophical papers*. New York: Oxford University Press.

Lewis, D K 1994 Humean supervenience debugged. *Mind*, 103(412): 473–490. DOI: https://doi.org/10.1093/mind/103.412.473

Lewis, D K 2000 Causation as influence. *Journal of Philosophy*, 97(4): 182–197.

Libet, B 1985 Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4): 529–566. DOI: https://doi.org/10.1017/s0140525x00044903

Libet, B 1999 Do we have free will? *Journal of Consciousness Studies*, 6(8–9): 47–57.

Libet, B 2002 The timing of mental events: Libet's experimental findings and their implications. *Consciousness and Cognition*, 11(2): 291–299. DOI: https://doi.org/10.1006/ccog.2002.0568

List, C 2014 Free will, determinism, and the possibility of doing otherwise, *Noûs*, 48(1): 156–178. DOI: https://doi.org/10.1111/nous.12019

List, C and Menzies, P 2009 Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy*, 106(9): 475–502. DOI: https://doi.org/10.5840/jphil2009106936

Löhrer, G and Sehon, S 2016 The Davidsonian challenge to the non-causalist. *American Philosophical Quarterly*, 53(1): 85–96.

Lombrozo, T and Carey, S 2006 Functional explanation and the function of explanation. *Cognition*, 99(2): 167–204. DOI: https://doi.org/10.1016/j.cognition.2004.12.009

Lowe, E J 2000 Causal closure principles and emergentism. *Philosophy*, 75(294): 571–586. DOI: https://doi.org/10.1017/s003181910000067x

Lowe, E J 2005 *The four-category ontology: A metaphysical foundation for natural science*. Oxford: Clarendon Press.

Lowe, E J 2008 *Personal agency: The metaphysics of mind and action*. New York: Oxford University Press.

Lowe, E J 2013 Substance causation, powers and Humean agency. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 153–173.

Lynch, M P and Glasgow, J 2003 The impossibility of superdupervenience. *Philosophical Studies*, 113(3): 201–221. DOI: https://doi.org/10.1023/a:1024037729994

Mackie, J L 1965 Causes and conditions. *American Philosophical Quarterly*, 2(4): 245–264.

Mackie, J L 1974 *The cement of the universe*. Oxford: Clarendon Press.

Malle, B 1999 How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1): 23–48. https://doi.org/10.1207/s15327957pspr0301_2

Malle, B 2004 *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B, Knobe, J and Nelson, S 2007 Actor–observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, 9(4): 491–514.

Mayr, E 2011 *Understanding human agency*. New York: Oxford University Press.

McDermott, M 2002 Causation: Influence versus sufficiency. *The Journal of Philosophy*, 99(2): 84–101. DOI: https://doi.org/10.5840/JPHIL200299219

McDonnell, N 2015 The deviance in deviant causal chains. *Thought: A Journal of Philosophy*, 4(2): 162–170. DOI: https://doi.org/10.1002/tht3.169

McDowell, J 1996 *Mind and world*. Cambridge, MA: Harvard University Press.

McKitrick, J 2005 Are dispositions causally relevant? *Synthese*, 144(3): 357–371. DOI: https://doi.org/10.1007/s11229-005-5868-z

McKitrick, J 2013 Getting Causes from Powers by Stephen Mumford and Rani Lill Anjum. *Analysis*, 73(2): 402–404. DOI: https://doi.org/10.1093/analys/ant016

Melden, A I 1961 *Free action: Studies in philosophical psychology*. London: Routledge & Kegan Paul.

Mele, A 1992a *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.

Mele, A 1992b Recent work on intentional action. *American Philosophical Quarterly*, 29(3): 199–217.

Mele, A 2000 Goal-directed action: Teleological explanations, causal theories, and deviance. *Noûs*, 34(14): 279–300. DOI: https://doi.org/10.1111/0029-4624.34.s14.15

Mele, A 2003 *Motivation and agency*. Oxford: Oxford University Press.

Mele, A 2005 Action. In: Jackson, F and Smith, M *The Oxford handbook of contemporary philosophy*. Oxford: Oxford University Press. pp. 78–88.

Mellor, D H 1995 *The facts of causation*. London: Routledge.

Melnyk, A 2003 *A physicalist manifesto: Thoroughly modern materialism*. New York: Cambridge University Press.

Melnyk, A 2006 Realization and the formulation of physicalism. *Philosophical Studies*, 131(1): 127–155. DOI: https://doi.org/10.1007/s11098-005-5986-y

Melnyk, A 2018 In defense of a realization formulation of physicalism. *Topoi*, 37(3): 483–493. DOI: https://doi.org/10.1007/s11245-016-9404-1

Menzies, P 2013 Mental causation in the physical world. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 58–88.

Menzies, P and Price, H 1993 Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44(2): 187–203. DOI: https://doi.org/10.1093/bjps/44.2.187

Mill, J S 1843 *A system of logic, ratiocinative and inductive being a connected view of the principles of evidence and the methods of scientific investigation*. London: Longmans, Green, Reader, and Dyer.

Millican, P 2007 Against the new Hume. In: Read, R and Richman, K *The new Hume debate: Revised edition*. London: Routledge. pp. 211–252.

Monroe, A and Malle, B 2017 Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, 146(1): 123–133. DOI: https://doi.org/10.1037/xge0000234

Morris, K 2018 Truthmaking and the mysteries of emergence. In Vintiadis, E and Mekios, C *Brute facts*. Oxford: Oxford University Press. pp. 113–129.

Mourelatos, A 1978 Events, processes and states. *Linguistics and Philosophy*, 2(3): 415–434. DOI: https://doi.org/10.1007/bf00149015

Mulligan, K, Simons, P and Smith, B 1984 Truth-makers. *Philosophy and Phenomenological Research*, 44(3): 287–321. DOI: https://doi.org/10.2307/2107686

Mumford, S 2004 *Laws in nature*. London: Routledge.

Mumford, S 2009 Passing powers around. *The Monist*, 92(1): 94–111. DOI: https://doi.org/10.5840/monist20099215

Mumford, S and Anjum, R L 2010 A powerful theory of causation. In: Marmodoro, A, *The metaphysics of powers: their grounding and their manifestations*. New York: Routledge. pp. 143–159.

Mumford, S and Anjum, R L 2011 *Getting causes from powers*. New York: Oxford University Press.

Mumford, S and Anjum, R L 2013 Causes as powers: Book symposium on Stephen Mumford and Rani Lill Anjum: Getting Causes from Powers. *Metascience*, 22(3): 545–559. DOI: https://doi.org/10.1007/s11016-013-9783-5

Nagel, T 1986 *The view from nowhere*. New York: Oxford University Press.

Ney, A 2008 Defining physicalism. *Philosophy Compass*, 3(5): 1033–1048. DOI: https://doi.org/10.1111/j.1747-9991.2008.00163.x

Noordhof, P 1999 The overdetermination argument versus the cause-and-essence principle—no contest. *Mind*, 8(430): 367–375. DOI: https://doi.org/10.1093/mind/108.430.367

O'Connor, T 2000 *Persons and causes: The metaphysics of free will*. New York: Oxford University Press.

O'Connor, T 2009 Agent-causal power. In: Handfield, T *Dispositions and causes*. Oxford: Oxford University Press.

O'Shaughnessy, B 1980 *The will*. Cambridge: Cambridge University Press.

Ott, W 2009 *Causation and laws of nature in early modern philosophy*. Oxford: Oxford University Press.

Papineau, D 1993 *Philosophical naturalism*. Oxford: Blackwell.

Papineau, D 2001 The rise of physicalism. In: Gillett, C and Barry, L *Physicalism and its discontents*. Cambridge: Cambridge University Press. pp. 3–26.

Papineau, D 2002 *Thinking about consciousness*. New York: Oxford University Press.

Paul, S 2020 *Philosophy of action: A contemporary introduction*. London: Routledge.

Peacocke, C 1979 Deviant causal chains. *Midwest Studies in Philosophy*, 4(1): 123–155. DOI: https://doi.org/10.1111/j.1475-4975.1979.tb00375.x

Pearl, J 2000 *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Pereboom, D 2014 *Free will, agency, and meaning in life*. New York: Oxford University Press.

Pettit, P 1993 A definition of physicalism. *Analysis*, 53(4): 213–223. DOI: https://doi.org/10.1093/analys/53.4.213

Prior, E, Pargetter, R and Jackson, F 1982 Three theses about dispositions. *American Philosophical Quarterly*, 19(3): 251–257.

Psillos, S 2002 *Causation and explanation*. London: Routledge.

Putnam, H 1975 The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7: 131–193.

Raley, Y 2007 The facticity of explanation and its consequences. *International Studies in the Philosophy of Science*, 21(2): 123–135. DOI: https://doi.org/10.1080/02698590701498035

Reid, T 1788 *Essays on the active powers of man*. Edinburgh: John Bell, Parliament-Square, and London: G G J & J Robinson.

Ryle, G 1949 *The concept of mind*. London: Hutchinson's University Library.

Salmon, W C 1984 *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Sartorio, C 2005 Causes as difference-makers. *Philosophical Studies*, 123(1/2): 71–96. DOI: https://doi.org/10.1007/s11098-004-5217-y

Schaffer, J 2007 Causation and laws of nature: Reductionism. In: Sider, T, Hawthorn, J and Zimmerman, D W *Contemporary debates in metaphysics*. Malden, MA: Blackwell, pp. 82–107.

Schaffer, J 2016 The metaphysics of causation. The Stanford Encyclopedia of Philosophy, 5 July 2016. Available at https://plato.stanford.edu/archives/spr2022/entries/causation-metaphysics [Last accessed 2 September 2023].

Schult, C and Wellman, H 1997 Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition*, 62(3): 291–324. DOI: https://doi.org/10.1002/icd.548

Searle, J 1983 *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.

Sehon, S 1997 Deviant causal chains and the irreducibility of teleological explanation. *Pacific Philosophical Quarterly*, 78(2): 195–213. DOI: https://doi.org/10.1111/1468-0114.00035

Sehon, S 2005 *Teleological realism: Mind, agency, and explanation*. Cambridge, MA: Bradford Book/MIT Press.

Sehon, S 2007 Goal-directed action and teleological explanation. In: Campbell, J K, O'Rourke, M and Silverstein, H S *Causation and explanation*. Cambridge, MA: MIT Press. pp. 155–170.

Sehon, S 2010 Teleological explanation. In: O'Connor, T and Sandis, C *A companion to the philosophy of action*. Oxford: Wiley-Blackwell. pp. 121–128.

Shapiro, L (ed.) 2007 *The correspondence between Princess Elisabeth of Bohemia and René Descartes*. Chicago, IL: University of Chicago Press.

Shapiro, S 1997 *Philosophy of mathematics: Structure and ontology*. New York: Oxford University Press.

Shepherd, J 2021 *The shape of agency: Control, action, skill, knowledge*. Oxford: Oxford University Press.

Shoemaker, S 1980 Causality and properties. In: van Inwagen, P *Time and cause*. Dordrecht: D. Reidel. pp. 109–135.

Shoemaker, S 2001 Realization and mental causation. In: Gillett, C and Loewer, *The proceedings of the Twentieth World Congress of Philosophy*. Cambridge: Cambridge University Press. pp. 23–33.

Shoemaker, S 2007 *Physical realization*. Oxford: Oxford University Press.
Shoemaker, S 2013 Physical realization without preemption. In: Gibb, S C, Lowe, E J and Ingthorsson, R *Mental causation and ontology*. Oxford: Oxford University Press. pp. 35–57.
Skillen, A 1984 Mind and matter: A problem which refuses dissolution. *Mind*, 93(372): 514–526. DOI: https://doi.org/10.1093/mind/xciii.372.514
Skow, B 2013 Are there non-causal explanations (of particular events)? *British Journal for the Philosophy of Science*, 65(3): 445–457. DOI: https://doi.org/10.1093/bjps/axs047
Skow, B 2018 *Causation, explanation, and the metaphysics of aspect*. Oxford: Oxford University Press.
Skyrms, B 1984 EPR: Lessons for metaphysics. *Midwest Studies in Philosophy*, 9(1): 245–255. DOI: https://doi.org/10.1111/j.1475-4975.1984.tb00062.x
Smith, M 2004 The structure of orthonomy. *Royal Institute of Philosophy Supplement*, 55: 165–193. DOI: https://doi.org/10.1017/s1358246100008675
Smith, M 2010 The standard story of action: An exchange 1. In: Buckareff, A A and Aguilar, J H *Actions: New perspectives on the causal theory of action*. Cambridge, MA: MIT Press. pp. 45–56.
Smith, M 2012 Four objections to the standard story of action (and four replies). *Philosophical Issues*, 22(1): 387–401. DOI: https://doi.org/10.1111/j.1533-6077.2012.00236.x
Smith, M 2021 Are actions bodily movements? *Philosophical Explorations*, 24(3): 394–407. DOI: https://doi.org/10.1080/13869795.2021.1957205
Steward, H 1997 *The ontology of mind: Events, processes and states*. Oxford: Oxford University Press.
Steward, H 2009a Sub-intentional actions and the over-mentalization of agency. In: Sandis, C *New essays on the explanation of action*. Basingstoke: Palgrave Macmillan.
Steward, H 2009b Animal agency. *Inquiry*, 52(3): 217–231. DOI: https://doi.org/10.1080/00201740902917119
Steward, H 2011 Perception and the ontology of causation. In: Roessler, J, Lerman H and Eilan, N *Perception, causation, and objectivity*. Oxford: Oxford University Press.
Steward, H 2012 *A metaphysics for freedom*. Oxford: Oxford University Press.
Steward, H 2013a Responses. *Inquiry: An Interdisciplinary Journal of Philosophy*, 56(6): 681–706. DOI: https://doi.org/10.1080/0020174x.2013.841055
Steward, H 2013b Processes, continuants, and individuals. *Mind*, 122(487): 781–812. DOI: https://doi.org/10.1093/mind/fzt080
Stoecker, R 2009 Why animals can't act, *Inquiry*, 52(3): 255-271. DOI: https://doi.org/10.1080/00201740902917135
Stout, R 2010 What are you causing in acting? In: Aguilar, J H and Buckareff, A A *Causing human actions: New perspectives on the causal theory of action*. Cambridge, MA: MIT Press.

Strawson, G 1987 Realism and causation. *The Philosophical Quarterly*, 37(148): 253–277. DOI: https://doi.org/10.2307/2220397

Strawson, G 1989 *The secret connexion: Causation, realism, and David Hume*. Oxford: Oxford University Press.

Strawson, P F 1985 Causation and explanation. In Bruce Vermazen, B and Hintikka, M B *Essays on Davidson: Actions and Events*. Oxford: Oxford University Press. pp. 115–135.

Stroud, B 1986 The physical world. *Proceedings of the Aristotelian Society*, 87(1): 263–277. DOI: https://doi.org/10.1093/aristotelian/87.1.263

Sturgeon, S 1998 Physicalism and overdetermination. *Mind*, 107(426): 411–432. DOI: https://doi.org/10.1093/mind/107.426.411

Tanney, J 1995 Why reasons may not be causes. *Mind and Language*, 10(1–2): 103–126. DOI: https://doi.org/10.1111/j.1468-0017.1995.tb00007.x

Tanney, J 2009 Reasons as non-causal, context-placing explanations. In: Sandis, C *New essays on the explanation of action*. Basingstoke: Palgrave Macmillan. pp. 94–111.

Tanney, J 2013 Ryle's conceptual cartography. In: Reck, E H *The historical turn in analytic philosophy*. Basingstoke: Palgrave Macmillan.

Taylor, R 1966 *Action and purpose*. Englewood Cliffs, NJ: Prentice-Hall.

Thompson, M 2008 *Life and action: Elementary structures of practice and practical thought*. Cambridge, MA: Harvard University Press.

Tooley, M 1990a Causation: Reductionism versus realism. *Philosophy and Phenomenological Research*, 50: 215–236. DOI: https://doi.org/10.2307/2108040

Tooley, M 1990b The nature of causation: A singularist account. *Canadian Journal of Philosophy*, 20(1): 271–322. DOI: https://doi.org/10.1080/00455091.1990.10717229

Turri, J 2018 Exceptionalist naturalism: Human agency and the causal order. *Quarterly Journal of Experimental Psychology (Hove)*, 71(2): 96–410. DOI: https://doi.org/10.1080/17470218.2016.1251472

Tye, M 1999 Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind*, 108(432): 705–725. DOI: https://doi.org/10.1093/mind/108.432.705

Van Fraassen, B C 1980 *The scientific image*. Oxford: Clarendon Press.

Velleman, D J 1992 What happens when someone acts? *Mind*, 101(403): 461–481. DOI: https://doi.org/10.7591/9781501721564-008

Vermazen, B 1985 Negative acts. In: Vermazen B and Hintikka, M B *Essays on Davidson: Actions and events*. Oxford: Clarendon Press. pp. 93–104.

von Wright, G H 1962 On promises. *Theoria*, 28(3): 277–297.

von Wright, G H 1971 *Explanation and understanding*. Ithaca, NY: Cornell University Press.

von Wright, G H 1974 *Causality and determinism*. New York: Columbia University Press.

Wang, Z 2013 *Process and pluralism: Chinese thought on the harmony of diversity*. Berlin: Walter de Gruyter.

White, A 2020 Processes and the philosophy of action. *Philosophical Explorations*, 23(2): 112–129. DOI: https://doi.org/10.1080/13869795.2020.1753801

Whittle, A 2008 A functionalist theory of properties. *Philosophy and Phenomenological Research*, 77(1): 59–82. DOI: https://doi.org/10.1111/j.1933-1592.2008.00176.x

Wilson, J 2005 Supervenience-based formulations of physicalism. *Noûs*, 39(3): 426–459. DOI: https://doi.org/10.1111/j.0029-4624.2005.00508.x

Wilson, J 2006 On characterizing the physical. *Philosophical Studies*, 131(1): 61–99. DOI: https://doi.org/10.1007/s11098-006-5984-8

Wilson, J 2010 What is Hume's dictum, and why believe it? *Philosophy and Phenomenological Research*, 80(3): 595–637. DOI: https://doi.org/10.1111/j.1933-1592.2010.00342.x

Wilson, J 2016 Grounding-based formulations of physicalism. *Topoi*, 37(3): 495–512. DOI: https://doi.org/10.1007/s11245-016-9435-7

Wittgenstein, L 1958 *The blue and brown books*. Oxford: Blackwell.

Woodfield, A 1976 *Teleology*. Cambridge: Cambridge University Press.

Woodward, J 2003 *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Yablo, S 1992 Mental causation. *Philosophical Review*, 101(2): 245–280. DOI: https://doi.org/10.2307/2185535

Yablo, S 1993 Is conceivability a guide to possibility? *Philosophy and Phenomenological Research*, 53(1): 1–42. DOI: https://doi.org/10.2307/2108052

Zagzebski, L 1994 The inescapability of Gettier problems. *Philosophical Quarterly*, 44(174): 65–73. DOI: https://doi.org/10.2307/2220147

Index

R

In *Understanding Mental Causation*, Andrea White proposes a compelling new approach to the problem of mental causation. Believing that contemporary philosophy of mind misunderstands mental causation, White explains where the leading theories go astray and offers a new theory, a radical departure from physicalism that solves critical problems for philosophers of mind and action.

Mental causation is often presented as a cause-effect relation between mental items and physical events. This relational understanding of mental causation seems to offer a straightforward explanation of what is going on when people act intentionally, but White argues it reduces intentional action to chains of causally related events, excluding the very thing we want to preserve – the agent. It has also prevented us from exploring more diverse accounts of the relationship between our mind and body, leaving physicalism as the dominant metaphysics of mind.

Instead of allowing ourselves to become trapped in a 'physicalist triad', White presents her own non-relational theory. Denying causation is always a relation, she holds instead that causation is a general type of process in which substances engage. She supports this view with a novel account of what processes are.

White shows how this theory can be used to provide a better understanding of intentional action and the mental causation associated with it. She suggests that to act intentionally is to engage in a process and, as such, to exercise a power – but a power of a special sort. To act intentionally is to wield a power to structure one's own activities so that they demonstrate a pattern. We then make sense of this pattern by appealing to mental concepts.

By reframing mental causation, *Understanding Mental Causation* offers a fresh starting point for developing theories of the mind and for asking new questions about action, mental causation and the mind-body connection.

WHITE ROSE
UNIVERSITY PRESS
Universities of Leeds, Sheffield & York

ISBN 978-1-912482-52-8
90000
9 781912 482528