REASONS, ACTIONS, AND CAUSES

by

CHIWOOK WON

A Dissertation submitted to the Department of Philosophy in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

BROWN UNIVERSITY

May 2013
This dissertation by Chiwook Won is accepted in its present form by the Department of Philosophy as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date ________________ ____________________________________

Jaegwon Kim, Advisor

Recommended to the Graduate Council

Date ________________ ____________________________________

James Dreier, Reader

Date ________________ ____________________________________

Richard G. Heck, Jr., Reader

Approved by the Graduate Council

Date ________________ ____________________________________

Peter Weber, Dean of the Graduate School
Abstract

Central to our conception of ourselves as agents is the idea that we act for reasons; we do things because of what we want, believe, hope, fear, and so forth. The overarching theme of this dissertation is the exploration of two ideas that are built into the concept of acting for a reason: the idea of rationality and the idea of cause. The dissertation falls into two parts. The first two chapters concern the questions of what reasons are and how they explain actions. The last two chapters concern the problem of mental causation, namely, how our actions can be a product of our mental states.

Chapter 1, “What It Is to Be a Reason” develops a general schema for analyzing our talk of reasons. On this schema, to say that there is a reason for an agent to $\phi$ is to say that there is a piece of reasoning leading to the agent’s $\phi$-ing that meets certain conditions. I characterize these conditions in terms of the agent’s proper belief and desire sets. I then argue that my analysis not only accords well with our ordinary ways of thinking about reasons, but also provides a useful framework for theorizing about reasons. In particular, I show how it can be used in formulating and clarifying various meta-normative distinctions, questions, and theses about reasons.

Chapter 2, “Action Explanation, Rationality, and Mental Simulation,” sketches a simulation-based account of action explanation. My approach is a normative one in that it recognizes an important role for normativity in analyzing the explanatory force of action explanation. But it is also ‘Davidsonian’ in that it can be seen as a simulationist reconstruction of Donald Davidson’s views on action explanation and rationality. On this account, briefly, reasons are rational (or rationalizing) causes in which ‘rational’ is to be understood in terms of being simulatable. I defend two main claims about this notion of rationality: First, rationality is the constitutive normativity of reasons, the normativity in
virtue of which an attitude counts as a reason for an action. Second, rationality sets limits on the actions or attitudes that we can understand, such that we can only understand the actions of those agents who share the same principles of rationality with us.

Chapter 3, “Overdetermination, Counterfactuals, and Mental Causation,” addresses the problem of mental causation, focusing on the so-called exclusion, or overdetermination, problem. In particular, I take up the question whether counterfactuals can yield an appropriate notion of causal redundancy, and examine how this issue bears on the mental causation debate. It is often suggested that the overdetermination argument relies on a certain ‘thick’ conception of causation, and that if causation were understood in terms of counterfactuals, the overdetermination problem simply would not arise. I try to motivate and give substance to this idea and then show why this strategy, though initially appealing, leaves the problem unresolved; rather, it only reveals the expressive poverty of counterfactual causal talk.

Chapter 4, “Mental Causation as Joint Causation,” explores the view that mental and physical events jointly cause physical events. This kind of view has usually been taken as a non-starter. It is widely assumed that although the view easily escapes the exclusion problem, it does so only at the high price of violating physical closure. As I show, however, this is doubly mistaken. First, once we get clear on two sources of the exclusion problem, physical closure and supervenience, we can see that the supervenience problem still arises for the joint causation view. Second, the view is consistent with a moderately strong principle of causal closure that respects much of our physicalist intuitions. I discuss how we can motivate and make sense of the view within an emergentist picture, and compare it with a prominent nonreductive physicalist conception of mental causation, the overdetermination view. Specifically, I distinguish between the overdetermination problem and the supervenience problem, and argue that the joint causation view fares better than the overdetermination view vis-à-vis both problems.
Acknowledgements

My greatest debt is to my advisor, Jaegwon Kim. Many of the ideas driving this work were born while I was taking his seminars and developed through his continued, invaluable feedback over the years. I am very grateful for his support and encouragement throughout my graduate study, and for being an example of the kind of philosopher I would like to be.

I am also grateful to the other members of my committee, Jamie Dreier and Richard Heck. Their thoughtful criticism and advice have strongly influenced my work, and my thinking on philosophical matters both connected and unconnected with this project. I consider myself very fortunate to be able to count them as my teachers.

I am also indebted to a great number of other philosophers at Brown. I would particularly like to thank Douglas Kutach and Kevin Morris, who provided helpful written comments on Chapters 3 and 4 respectively. Each Chapter of the dissertation has benefited from presentations at the Dissertation Workshop. Among the participants of the Workshop, Derek Bowman and Eoin Ryan require special mention for their detailed comments and discussion on my dissertation at various stages.

I would also like to take the opportunity to thank all my teachers in Korea. Chapter 3.2 is descended from a chapter of my 2006 MA thesis at Seoul National University, which benefited from criticism and encouragement, specially from Sungho Choi and Inkyo Chung. I am also grateful to my MA advisor, Young-Jung Kim, who guided me in the last years of his life.

Finally, I would like to thank my family and particularly my parents for their love, encouragement, and unshakeable faith in me. My deepest personal thanks to my wife and daughters for making clear what is most important in life.
Contents

Abstract v

Chapter 1. What It Is to Be a Reason 1
   1. Introduction 1
   2. Reasons and Reasoning 3
   3. Humean Theories of Reasons 8
   4. Externalist Humeanism and Agent-neutral Reasons 11
   5. Reasons Ascriptions 16
   6. Some Metaphysical Issues 20

Chapter 2. Action Explanation, Rationality, and Mental Simulation 22
   1. Introduction 22
   2. The Dray-Hempel Debate 24
   3. Reasons as a Species of Cause: Making Dray Davidsonian 27
   4. Gordon and Kim on Action Explanation and Simulation 30
   5. The Constitutive Normativity of Reasons: Rationality 37
   6. Rationality as Simulatability: Making Davidson Simulationist 41
   7. Circles of Rational Intelligibility 44
   8. Conclusion 47

Chapter 3. Overdetermination, Counterfactuals, and Mental Causation 49
   1. Introduction 49
   2. The Overdetermination Problem 51
   3. Overdetermination and Counterfactuals 56
   3.1. Counterfactual Analyses of Overdetermination 57
CHAPTER 1

What It Is to Be a Reason

In this chapter, I develop a general schema for analyzing our *reasons* talk. The central claim, simply put, is that statements about reasons are best understood as statements about *reasoning*. More specifically, on my proposed analysis, to say that there is a reason for an agent to φ is to say that there is a piece of practical reasoning leading to the agent’s φ-ing that meets certain conditions. I characterize these conditions in terms of the agent’s *proper* belief and desire sets. I then argue that my analysis not only accords well with our ordinary ways of thinking about reasons, but also provides a useful framework for theorizing about reasons. In particular, I show how it can be used in formulating and clarifying various meta-normative distinctions, questions, and theses about reasons.

1. Introduction

Reasons are everywhere in our cognitive and practical life. When we try to understand why a person acted as she did, we try to see her reason for acting that way. When we evaluate her action, we assess whether there is a good or justified reason for her to act in that way. And when we deliberate about what to do, we reflect on reasons for, or against, acting in certain ways, trying to give due weights to them. Thus, reasons, in particular practical reasons, are what we deliberate about and what explain and justify our actions. It is no surprise that the notion of a reason is central to much philosophical discussion in moral theory, philosophy of mind and action, and other related areas.

But what are we saying when we make a claim about reasons, for example, that someone has a reason to do something? Despite the familiarity and centrality of the notion of a reason in everyday and theoretical contexts, there is no general agreement, even a rough one, over what reason statements mean or what reasons are. Sometimes the impression one gets from looking into the literature is that it is not clear even whether participants in
debates on reasons are talking about one and the same thing. To take the familiar example: You falsely believe that the glass in front of you contains gin; you drink it, but it was actually petrol (Williams 1979). In this case, on the one hand, it sounds natural to say that there was no reason for you to drink the stuff; maybe you thought you had one, but you didn’t. On the other hand, it is also perfectly natural to say that your reason to drink it was that you believed it was gin. It seems then that we are contradicting ourselves in making these two claims in the same breath.

Observations of this sort lead many to hold that there are different senses or kinds of reasons for action. It is often said that when we claim that you have no reason to drink it, we are talking about normative reasons, while we are talking about motivating reasons when we claim that your reason for drinking it was your belief that it was gin. The distinction between normative and motivating reasons is now widely accepted. It is not clear, however, exactly how the distinction is to be made, nor what exactly it amounts to. Does it mean that the notion of a reason is not univocal, like ‘bank’? Or are there actually two species of reasons? How then are they related to each other? In fact, a number of other distinctions have also been drawn—to name a few, distinctions between internal and external reasons, objective and subjective reasons, agent-neutral and agent-relative reasons, and so on. And we now see these distinctions being commonly used with little apparent regard for whether their use is underpinned by a consistent, tolerably unified, and shared meaning of reason.

A related concern is with what might be called the ontology of reasons. What are we talking about when we talk about reasons? Usually, philosophers who are primarily concerned with the explanatory role of reasons tend to take reasons to be psychological states, while those who are primarily concerned with normative or evaluative aspects of reasons tend to take them to be facts or propositions. However, various other kinds of things are also said to be reasons—for example, considerations, states of affairs, events, goals, aims, and so on. There is no consensus even about what sort of thing, ontologically, reasons are.
All this, I take it, reflects the fact that the concept of reason plays a variety of roles in our practical life. But this need not mean that reasons are “strange beasts,” as some philosophers put it (Kearns and Star 2009, 215). Nor does it mean that we cannot have a unifying and systematic understanding of reasons. The goal of this chapter is to give a sketch of such an account, a coherent and unified account of reasons in which various uses of ‘reason’ turn out to make sense and multiple roles of reasons in deliberating, explaining, and evaluating actions are accommodated. Specifically, what I will try to do is to develop a general schema for analyzing, and systematizing, our talk of reasons. My assumption here is that reasons of any kind (or sense) have something in common in virtue of which they all count as reasons. More specifically, I believe that claims about reasons are best understood as claims about reasoning. Crudely, to say that someone has a reason to $\phi$ is to say that there is a piece of reasoning that leads to the agent’s $\phi$-ing. This might sound obviously true or even trivial—indeed, I hope it does. This chapter is an attempt to systematically explore this simple idea. In what follows, I show that my account is not only adequate as an analysis of our ordinary concept of reason, but also provides a useful framework for theorizing about practical reasons—a framework for reason discourse that helps formulate and clarify various meta-normative distinctions, questions, and theses about reasons. I close by briefly discussing the metaphysical upshot of my account as compared with T. M. Scanlon’s Reasons Fundamentalism.

2. Reasons and Reasoning

Some philosophers deny that it is possible to give an interesting analysis of the concept of a reason. Scanlon (1998), for example, thinks that there being a reason for an action is just there being a consideration that “counts in favor of” the action, and that any attempt to explain how a reason counts in favor of an action will never be able to go far. According to Scanlon, “‘By providing a reason for it’ seems to be the only answer” (1998, 17). Similarly, Derek Parfit claims that like some other fundamental concepts, “the concept of a reason is indefinable in the sense that it cannot be helpfully explained merely by using words” (2011,

$^1$I believe my account can be extended to epistemic reasons, but here I focus on practical reasons.
31). I think otherwise. I believe there is a lot more we can say about how a reason favors an action than Parfit and Scanlon suppose.

It seems that the concept of reasons is closely connected with the concept of reasoning. What is it to say that, for example, there is a reason for Mary to take an umbrella with her? As a first pass, we might hold, it is to say that there is a piece of practical reasoning leading to her taking the umbrella, like this:

(b) Mary will get wet unless she take an umbrella;  
(d) Mary doesn’t want to get wet;  
(φ) So, Mary takes an umbrella.

This is only a first approximation, and it raises many questions. First, what is reasoning or practical reasoning? Like the term ‘reason,’ ‘reasoning’ is used in many different ways. Let’s focus for now on a somewhat primitive sense of reasoning in the following sense: reasoning is a process in which a subject moves from one set of intentional states to others that are appropriately related to it. A little more precisely, and also stipulatively, I will represent practical reasoning as a sequence of items, \( \langle a, b, c, \ldots, \phi \rangle \), where the antecedent items, certain attitudes like beliefs and desires, stand in appropriate normative relation to \( \phi \), an intention or course of action.

What is meant by the ‘appropriate normative relation’? It is meant to capture the idea that for a given sequence of items to constitute a piece of reasoning, the items must be suitably related to each other, in virtue of which relation the sequence counts as a piece of reasoning. For instance, it seems that Mary’s belief and desire about rain in the above reasoning clearly bear a certain sort of normative relation to her taking an umbrella. For comparison, imagine this sequence: \( \langle \text{I believe snow is white; I want to get wet; so I take an umbrella} \rangle \). This seems to be just a certain array of content-bearing entities. Unlike this, \( b \) and \( d \) in Mary’s reasoning stand in some sort of normative relation, i.e., the means-end relation, to \( \varphi \). That’s why the sequence \( \langle b, d, \varphi \rangle \) deserves to be called reasoning.
This sort of instrumental reasoning is perhaps the simplest form of practical reasoning. It is arguable that the means-end relation is the most fundamental kind of normative relation that underwrites practical reasoning. One might even want to say that the instrumental relation of means to ends is the only constitutive normative relation of practical reasoning. We need not settle these issues here. In any case, it is not part of my analysis that the means-end relation is the only appropriate normative relation that underwrites reasoning or reasons, nor that all practical reasoning is ultimately based on instrumental reasoning. There might be other sorts of basic, appropriate normative relations that can hold between an action and the antecedent items of reasoning. If there indeed are distinct normative relations that constitute practical reasoning, we can take them as defining distinct kinds of reasons. We can regard then the ‘appropriate normative relation’ as a placeholder for whatever normative relation underwrites practical reasoning. Let us use the term ‘R-relation’ for short to refer to the relation the premise of reasoning bears to its conclusion in virtue of which a given sequence counts as practical reasoning. We may then say that the set of antecedent items of reasoning is R-related to its concluding intention or action. R-relations, in short, are what makes a sequence of attitudes an instance of reasoning.

Now, how should we understand the condition that there is a piece of reasoning, say, leading to Mary’s taking the umbrella? Different ways of construing it will correspond to different senses of reasons. Let us start with perhaps the simplest idea: to say that someone has a reason to take an umbrella is to say that she actually performs a bit of practical reasoning that leads to her taking an umbrella, like \( \langle b, d, \phi \rangle \). This shouldn’t sound plausible, however, especially if it requires that the agent consciously go through the process of that reasoning. This is so even when we assume that we only talk about reasons that actually motivate the agent. It is implausible to suppose that whenever there is a reason that motivates an agent to act in a certain way, the agent consciously has engaged in a certain process of reasoning, just as it is implausible to suppose that whenever there is

---

2See, for example, Dreier (1997) for a discussion of why the means/endpoint rule has a kind of ground-level normative status.

3For example, why are practical reasons and epistemic reasons of distinct kinds? It’s because, on this way of thinking, practical reasoning and theoretical reasoning are regulated by different normative principles.
a reason for which an agent acts, the agent consciously engages in a process of reasoning—many of our actions are almost automatic.

However, it does seem plausible to suppose that an agent must have some relevant desires and beliefs to have such a reason, though it doesn’t require certain reasoning to have occurred in the agent’s consciousness. Suppose Mary actually takes an umbrella and, knowing this, we wonder why. In such a context, what matters is what the agent believes and desires. Only what Mary actually believes or desires can possibly be a reason that motivates, hence potentially explains, her action. When we say that Mary has a reason to $\phi$ in that context, we can be understood as saying something like this: there is a piece of reasoning leading to her $\phi$-ing, like $\langle b, d, \phi \rangle$, such that Mary actually has those attitudes $b$ and $d$.

This is the sense of reason that is salient in the context of explaining actions and what the term ‘motivating reason’ is intended to capture. When we use the notion of reason in this sense, having a reason requires having some relevant attitudes. Also, having some relevant attitudes seems to suffice for having a reason. If Mary does believe that she will get wet unless she takes an umbrella (and also wants not to get wet), she clearly has a reason to take an umbrella, whether she actually ends up taking an umbrella or not.

We can state the idea more explicitly. Let us first put the basic, canonical form of practical reasoning this way: A sequence of items, $\langle r, s, \phi \rangle$, is a piece of reasoning when $r$ is a representational state like belief, $s$ is a motivational state like desire, and they are $R$-related to $\phi$—that is, they stand in an appropriate normative relation, like the means-end relation, to $\phi$. Then, one thing you might mean when you say that there is a reason for someone to $\phi$ is this: There is a piece of reasoning, $\langle r, s, \phi \rangle$, such that $r$ and $s$ are among what the agent believes and desires.

This is not the only way we use the term reason, however. We might well want to say that Mary has a reason to take an umbrella even if she has no idea about the weather outside,
or even if she mistakenly believes it’s sunny outside, if she really dislikes getting wet. Or, even if she does not dislike getting wet, or even if she rather likes getting wet, Mary might still have a reason to take an umbrella—suppose, for example, getting wet will make her sick. In order to capture the full range of reason talk, then I introduce the following general schema:

**Schema R:** To say that there is a reason for agent A to φ is to say that there is a piece of reasoning, \(r, s, \phi\), such that \(r\) is in A’s *proper* belief set, \(R\), and \(s\) is in A’s *proper* desire set, \(S\).

On this schema, that someone has a reason to φ means that there is a piece of reasoning ending with φ whose premises are in the agent’s *proper* belief and desire sets, \(R\) and \(S\). Here, \(R\) and \(S\) are variables to be partly settled by the context. For example, in the context of explaining an action, or when our attention is on matters of motivation, the agent’s *actual* belief and desire sets will be appropriate for her \(R\) and \(S\) respectively. This would not be so, however, if we talk about reasons in a fully normative sense. There will be a context, for example, in which we might well say that Mary has a reason to take an umbrella, though she has no idea about the weather outside. On my approach, what we mean here is this: there is a piece of reasoning \(b, d, \phi\) where \(b\) is in Mary’s \(R\) and \(d\) is in her \(S\), though they are not what she actually believes or desires. Similarly, when we say that Mary has a reason to take an umbrella even though she likes getting wet, what we mean is this: there is a piece of reasoning like \(\text{I will get sick unless I take an umbrella; I don’t want to get sick; so I take an umbrella}\) such that the desire not to get sick is in Mary’s \(S\) and the means-end belief is in her \(R\). In general, in contexts of evaluating an action, we are not merely concerned with what the agent actually believes or desires. Intuitively, we can understand \(R\) and \(S\) as the set of *normative alternatives* to the agent’s actual belief and desire sets, something like the sets of beliefs and desires the agent ought to have.

There is room for a number of distinct ways of specifying \(R\) and \(S\). Different specifications of them will serve different evaluative purposes. Suppose Mary, who has no idea about the weather outside, hears her father saying that it’s raining outside, and so carries
an umbrella with her. But, suppose, he was mistaken, and it actually is not raining. In this case, on one hand, there is a sense in which Mary had a reason to take an umbrella; it was a reasonable thing to do, you might say, given the information or evidence available. On the other hand, there also is a sense in which there was no reason for her to take an umbrella; it wasn’t raining in fact, so the umbrella serves no purpose. This is the distinction between what we may call subjective and objective senses of reason. That is, there is a subjective reason, but no objective reason, for Mary to take an umbrella. Roughly, subjective reasons depend on the agent’s doxastic attitudes, while objective reasons depends on how things actually are. This distinction can be put within my framework this way: In some contexts, we will need to specify $R$ and $S$ in ways that are solely grounded in the agent’s psychology, while in some contexts, we need to also take her actual situation into account in specifying $R$ and $S$.

Now, it is the objective sense of reason that has come to occupy a central place in moral philosophy and metaethics—this is, I take it, what Scanlon (1998) calls the “standard normative sense” of reason. In what follows, I will be concerned with this sense of reason unless otherwise noted.

3. Humean Theories of Reasons

In the picture outlined above, what $R$ and $S$ are is usually settled by context. Sometimes, however, context is not enough to settle them. This happens, for example, when there is genuine normative disagreement. Suppose you and I are disputing over whether Mary has reason to take an umbrella. Mary doesn’t care about getting wet at all, but she has a cold. There is more than one way you and I can be disagreeing. One point we can disagree on is whether having a cold is a reason for Mary to take an umbrella. To put it within my framework, we can disagree on whether the belief (or fact or proposition) that Mary has a cold is in her proper $R$ and whether it, together with some appropriate element in her $S$, is $R$-related to her taking an umbrella. One thing this illustrates is that given the context, different ways of specifying $R$ and $S$ can be seen as generating, or corresponding to, different substantive theories of reasons—here, different substantive theories of objective normative
reasons—that is, normative theories which tell us what an agent has a reason to do on a given occasion.

For example, here is a naive, probably too naive, theory of reasons. Consider an agent who has a certain set of beliefs and desires in a given circumstance. We may want to ask what the agent has a reason to do. Let us assume that the means-end relation is the only appropriate $R$-relation that underwrites practical reasoning. The naive theory then says that the agent’s $R$ and $S$ are just the agent’s actual belief and desire sets, $B$ and $D$, respectively. According to Schema R, an agent has a reason to $\phi$ just in case there is a piece of reasoning, $\langle r, s, \phi \rangle$, in which $r$ and $s$ are in the agent’s $R$ and $S$. If $R$ and $S$ are just the agent’s $B$ and $D$, what emerges will be an account like this: an agent has a reason to $\phi$ iff the agent has some desire and believes that it will be served by her $\phi$-ing.

This is a Humean theory in that reasons are desire-dependent. On this naive Humean theory, what reasons one has is entirely determined by what she presently believes and desires; in fact, the distinction between motivating and normative reasons collapses. This perhaps is both too broad and too narrow. On the one hand, the theory seems to be too permissive because it says that Mary has a reason to take an umbrella when she mistakenly believes it’s raining. On the other hand, it seems to be too restrictive because it says that Mary has no reason to take an umbrella when she wants to keep dry, but doesn’t know it’s raining. The problem arises from the fact that an agent could have false beliefs, but we may not want to allow false beliefs to ground her reasons. This suggests a natural way to modify the naive theory. This would be to require that an agent’s proper belief set, $R$, be not her actual belief set, $B$, but something like her informed or corrected belief set, $B^*$. (Or you might want to make $B^*$ a set of relevant facts or true propositions.) Then what results will be an account like this: an agent has a reason to $\phi$ just in case the agent has some desire the satisfaction of which will be served by his $\phi$-ing, as a matter of fact, or so the informed agent would believe.

This account can accommodate the claim that Mary has a reason to take an umbrella when she doesn’t know it’s raining. It’s because there is the instance of practical reasoning,
\[\langle b, d, \varphi \rangle, \text{ in which } d \text{ is in her } S, \text{ i.e., } D, \text{ and } b \text{ is in her } R, \text{ i.e., her informed belief set } B^\ast. \] Similariy, the modified account would not be committed to saying that Mary has a reason to take an umbrella when she mistakenly believes it’s raining outside—\( b \) would not be in her \( R \), i.e., \( B^\ast \), in such a case. The modified account, however, might still be both too broad and too narrow. On the one hand, it might be thought too permissive in that any actual desire of the agent can give rise to a reason. Just as the agent’s actual belief may be incorrect, her actual desire may be intrinsically perverse or based on false beliefs, in which case you may want to say that it doesn’t provide a reason. On the other hand, it might seem too restricted in that only actual desires of the agent can generate reasons. If we are to allow changes or revisions of the agent’s beliefs in considering what reasons she has, why not allow changes or revisions in the agent’s desires? Indeed, our desires do and should change according to our beliefs, if we are rational (but not vice versa). It may then seem plausible to require that once we are allowed to revise or update the agent’s beliefs in considering what she has a reason to do, we are also allowed, or required, to update the agent’s desires accordingly. If so, a natural modification is to take the agent’s informed desire set, not her actual desire set, as her proper desire set. Then we have the following account: an agent has a reason to \( \phi \) iff there is a piece of reasoning, \( \langle r, s, \phi \rangle \), where \( r \) is in her informed belief set and \( s \) in her informed desire set.\(^5\)

On this kind of account, for an agent to have a reason to \( \phi \), it is not required that \( \phi \)-ing serve to promote the agent’s current desires; it is only required that \( \phi \)-ing would serve to promote the agent’s informed desires. Thus, what reasons one has are not determined by what the agent currently believes and desires. To make the account more substantive, the Humean would need to spell out how to understand an agent’s informed beliefs and desires. One thing we may naturally expect about the agent’s informed attitudes is that they should be somehow appropriately connected to, or based on, the agent’s actual attitudes. Indeed, it is this sort of connection that Bernard Williams insists on when he defends his internalism.

\[^5\text{There will be various nonequivalent ways of characterizing the agent’s informed belief or desire set, which will generate different versions of a Humean theory of reasons. Richard Brandt’s idea of “cognitive psychotherapy,” which is supposed to “clean up” the agent’s desires, gives a good example. See Brandt (1979). See also Smith (1994) where he appeals to “reflective equilibrium.”}\]
about reasons. His basic idea can be put in my framework this way: An agent’s proper desire set, \( S \), should be dependent on or “relativized to” her actual desire set \( D \)—more specifically, \( S \) should be reachable from \( D \) by a “sound deliberative route” (Williams 1979, 1989).

Broadly, this internalism/externalism debate is a debate over the conditions under which an agent has a reason to do something. Specifically, it concerns how those conditions are related to the agent’s inner conditions, particularly her motivations. To put it another way, it is a dispute over how facts about reasons are related to facts about the agent’s psychology, particularly, facts about her motivation. We can put the debate in my framework this way: it concerns how an agent’s proper desire set, \( S \), is related to her actual desire set \( D \). More generally, we can ask how an agent’s \( R \) and \( S \) should be related to her actual \( B \) and \( D \). Further, we may want to know what kind of general and plausible constraints we can impose on \( R \) and \( S \). As I will try to illustrate, this way of putting these questions helps us see exactly what is at stake in such debates over reasons.

4. Externalist Humeanism and Agent-neutral Reasons

Schema R says that an agent has a reason to \( \phi \) just in case there is a sequence of items, \( \langle r, s, \phi \rangle \), where \( r \) and \( s \) are in the agent’s \( R \) and \( S \) respectively and they are \( R \)-related to \( \phi \). On this schema what reasons an agent has is fully determined by the agent’s \( R \) and \( S \), given the \( R \)-relation. Suppose we want to know whether Mary has a reason to go to the party tonight. Then we need only to check Mary’s \( R \) and \( S \) to see if there is a pair of their members which are \( R \)-related to her going to the party. If Mary’s \( S \) contains, for example, a desire for gin and her \( R \) contains a belief or fact that she can get gin at the party, then she has a reason to go to the party. Thus, on Schema R, what an agent has a reason to do is determined by the agent’s \( R \) and \( S \).

Let us consider what general constraints could be given to an agent’s \( R \) and \( S \), and what possible views might result. We have seen how several Humean theories of reasons can be represented in my framework in terms of the specification of \( R \) and \( S \). Roughly put, the central idea of Humeanism is that what reasons one has depends on one’s psychology,
particularly on what one desires. This allows a number of nonequivalent interpretations. For instance, here is one loose way of construing the idea: If the agent’s desires were different, she would have different reasons. To put it in my framework, an agent’s \( \langle R, S \rangle \) should depend on the agent’s \( \langle B, D \rangle \). How should, then, an agent’s \( \langle R, S \rangle \) be related to her \( B \) and \( D \)? Does it depend solely on her \( B \) and \( D \), or are there other factors on which it depends? If it depends on the agent’s desires, does it mean that there can be no reasons that are not contingent on the agent’s psychology?

It seems plausible that different people have different reasons for action. First, they may have different reasons because what they want or believe is different. Mary, who desires gin, may have a reason to go to the party, but Billy who hates gin may not. Second, people may have different reasons because they are in different circumstances. Mary who lives in London may have a reason to take an umbrella with her when she goes out, but Billy who lives in Sydney might not. One might want to deny one (or both) of these sources of difference in reasons. One might hold, for example, that what reasons an agent has can be determined purely by the agent’s circumstances or environments. Or one might hold that reasons are determined purely by internal facts about the agent’s mental states. Imagine an agent in a possible world who has beliefs and desires exactly the same as yours. Could it be that this agent has a reason to \( \phi \) while you don’t have a reason to \( \phi \)? If one believes that facts about reasons are determined by or supervene on internal facts about the agent, the answer would be no. Let us call this view Internalism about reasons.\(^6\) According to Internalism, whether one has a reason to do something is always determined purely by the agent’s mental states. The naive Humean theory of reasons is a good example of Internalism. On this model, an agent’s \( R \) and \( S \) are just her actual \( B \) and \( D \), so what she has a reason to do depends entirely on what she believes and desires.

*Externalists* about reasons, as we may call them, deny this idea. Consider Mary who wants gin and believes that there is going to be gin at the party. Is there a reason for Mary

---

\(^6\)Obviously this is different from what Williams meant by his internalism. My terminology would sound more natural in relation to subjective reasons and epistemic reasons. See, e.g., Wedgwood (2002) for internalism about epistemic reasons (or rationality) in this sense.
to go to the party? Externalists say that it depends—it depends, for instance, on whether there really is going to be gin at the party. If there is, Mary has a reason; if not, not. Thus, according to Externalism, facts about reasons are not determined solely by the agent’s internal states, but also on something external to the agent, namely, features of the external world. Externalism, as I define it here, is just the denial of Internalism.

There are several different forms of Externalism. First, there might be Externalists who believe that what reasons an agent has depends solely on external facts, namely, the circumstances where the agent is. The thought would be that if you fix all the facts external to the agent, you thereby fix all the facts about the agent’s reasons, regardless of what the agent believes or desires. Let us call this view *Extreme Externalism* about reasons. We can formulate the idea this way:

**EE**: For any circumstance $C$, and for any agents $A_i$ and $A_j$, $A_i$ in $C$ has a reason to $\phi$ if and only if $A_j$ in $C$ has a reason to $\phi$.

**EE** says that if two agents were in the same circumstances, they would share every reason for action whatever their beliefs and desires are. This may sound too extreme to be plausible. I think it might turn out to be theoretically useful, or even plausible, depending on what kind of reasons we are concerned with—for example, when we restrict our attention to moral reasons.

Second, there may be Externalists who want to say that it will not be the case that agents in the same circumstances share every reason for action, but it may be true that there are reasons that commonly hold for any agents as long as they are in the same circumstances. Let us call this view *Moderate Externalism*. We can formulate it as follows:

**ME**: For some circumstance $C$, there is a $\phi$ such that any agents $A_i$ in $C$ has a reason to $\phi$.

**ME** says that for a certain circumstance, there are certain reason statements that are true of us whatever our mental states are. Suppose your sister Mary is starving to death. We may want to say that you have a reason to help Mary no matter what you believe and
desire. The idea is that an agent might have a reason only in virtue of certain facts about her circumstances.

Now, there might be Externalists who do not even endorse this moderate form of Externalism. Let us call them Timid Externalists. Timid Externalism rejects the idea that some of the agent’s reasons can be determined independently of her psychology. On this view, there is no guarantee that any two agents under the same circumstances would share some of their reasons. Still Timid Externalists are Externalists in that they reject the idea that reasons are determined purely by internal facts of the agent. On their view, no reason is either psychologically determined or entirely environment-determined.

To sum up, Internalism says that every reason is purely psychologically determined; Extreme Externalism says that every reason is purely environment-determined; Moderate Externalism says that some reasons are purely environment-determined; Timid Externalism says that no reason is purely environment-determined or psychologically determined.

We can put the ideas in terms of R and S in this way: Internalism is the view that an agent’s ⟨R, S⟩ is determined by, or supervenes on, the agent’s ⟨B, D⟩. Externalism denies this. In particular, Extreme Externalism holds that ⟨R, S⟩ depends only on the agent’s circumstance C. Moderate Externalism rejects this; it holds that ⟨R, S⟩ depends also on the agent’s psychology, so that ⟨R, S⟩ is a function not only of C but also of ⟨B, D⟩. Nonetheless Moderate Externalism accepts that there are some reasons that hold for us whatever our mental states are, which Timid Externalism denies.

Now, Humeans believe that reasons are grounded in the agent’s psychology, particularly in her motivations. Humeanism, as I characterized it, is the view that an agent’s ⟨R, S⟩ depends, wholly or partly, on the agent’s ⟨B, D⟩. Among Humeans, naive Humeans identify the agent’s ⟨R, S⟩ with her ⟨B, D⟩. So they are Internalists. Typically, however, popular versions of Humeanism take an Externalist form. Bernard Williams’ view is a good example. On such a view, an agent’s ⟨R, S⟩ does not supervene on the agent’s psychology; rather, it should be appropriately connected to ⟨B, D⟩ by a sound deliberative route. Thus, on Williams’ model, ⟨R, S⟩ depends both on ⟨B, D⟩ and C.
So Williams’ view is a form of Externalist Humeanism. One important, general question here is how far Humeans can go in the Externalist direction. Obviously, Humeanism is not compatible with Extreme Externalism; Humeans hold that reasons depend on desires. Then, can Humeans be Moderate Externalists? It might seem they cannot, for Moderate Externalism holds that there can be reasons that are determined independently of the agent’s psychology. If so, it might seem that Humeanism leads to a certain kind of skepticism about moral reasons, insofar as they are supposed to obtain independently of the agent’s psychology. Traditionally, that has been at the heart of concerns about Humeanism. If reasons are grounded in our motivations, which vary from person to person, how could there be a reason that holds for everyone regardless of differences in our psychology?

I believe that the basic idea of Humeanism is consistent with Moderate Externalism, so that answer to this question depends on the details of how you flesh out a Humean theory—for instance, on exactly how to understand Williams’ sound deliberative route. Moderate Externalism says that there are circumstances in which some of an agent’s reasons are determined independently of the agent’s psychology. Let us call such reasons agent-neutral reasons. Agent-neutral reasons need not be reasons that any agent whatsoever has, which we may call universal reasons. One may have an agent-neutral reason to φ in C1, while not in C2. But others would have that reason too if they were in C1, if the reason is indeed agent-neutral. Here is one way of putting the idea: There will be an agent-neutral reason to φ for a given circumstance C if there are r and s such that r is in R of any agent in C, s is in S of any agent in C, and they are R-related to φ.7

Consider an agent who has ⟨B1, D1⟩ in C1. Suppose a Humean theory assigns ⟨R1, S1⟩ to that agent. Then the Humean theory will tell us what the agent has a reason to do in his circumstance. Humeanism, as I loosely characterized it, is the view that the agent’s ⟨R, S⟩ depends on her ⟨B, D⟩. That is, holding fixed the circumstance, different ⟨B, D⟩

7 An agent-neutral reason also allows a weaker reading: There will be an agent-neutral reason to φ in C if there is φ such that for any agent A in C, there are r and s such that r is in A’s R, s is in A’s S, and they are R-related to φ. Several philosophers have drawn the distinction between agent-neutral/agent-relative reasons in different ways. See, e.g., Nagel (1970), Parfit (1984), Schroeder (2007). My usage of ‘agent-neutral’ is closest to Schroeder’s.
will generate different \( \langle R, S \rangle \). If the agent in \( C1 \) had not had \( \langle B1, D1 \rangle \) but had \( \langle B1, D2 \rangle \), for instance, the Humean theory would have assigned a different \( \langle R, S \rangle \) to the agent, so that agents with different desires come out to have different reasons. Thus understood, however, there is nothing in Humeanism to preclude agent-neutral reasons. There might be some reasons shared by any agents in \( C1 \) with different \( \langle B, D \rangle \). It could be the case, for example, although agents in \( C1 \) with different \( \langle B, D \rangle \) are assigned different \( \langle R, S \rangle \), all these different \( \langle R, S \rangle \) may have common elements. They will then be able to ground reasons that anyone in \( C1 \) has. In fact, Humeanism characterized above is consistent even with universal reasons. For example, it might be that there are some beliefs and desires that are constitutive of agency, so that they are arguably in everyone’s \( R \) and \( S \). They might then ground reasons that any agent must have.

5. Reasons Ascriptions

Thus far I have tried to illustrate how my analysis can provide a useful framework for reasons discourse that helps formulate and clarify various theoretical distinctions, questions, and theses about reasons. In this section, I argue that my account also fits well with the ways we ordinarily think and talk about reasons, and ascribe them to people.

Much of our talk about reasons is about so-called \( pro \ tant \) reasons. When we say that a person has a reason to \( \phi \), the reason must have some weight, but it can be outweighed by other reasons. So we allow that an agent has many reasons, reasons to do many—often conflicting—things. In addition, the agent might have multiple reasons for (or against) doing each of them. On the proposed approach, all this means is that there are pieces of \( reasoning \) for many actions, and there may also be multiple pieces of reasoning for (or against) each of the actions. Consider, for example, Mary who has a party tonight. She has reasons to go to the party; she likes gin, and Billy, who she loves, will be at the party. But she also has a reason not go to the party; she has a paper to finish tonight. Together this means that there are pieces of reasoning for and against Mary’s going to the party.

Thus, on my approach, claims about reasons are basically claims about reasoning. I said earlier a little about what I mean by reasoning here. When I use the term ‘reasoning’
or ‘practical reasoning,’ I am not invoking the ordinary notions like practical *deliberation* or whatever calculative activity you would engage in when you deal with your everyday decision-making problems. Rather, it is a technical term to capture the minimal, core sense of reasoning that I take to ground the ordinary sense. I don’t think that the notion of reasons can be understood in terms of such notions as deliberation or decision-making inferences. Instead, such notions *presuppose* reasons. Deliberation or real-life reasoning involves thinking about reasons. Basically it is a matter of weighing conflicting reasons for and against competing options. When Mary deliberates about what to do tonight, for example, she is deliberating about what reason (or reasoning) she should act on, trying to give due weight to the options and reasons. Reasons or reasoning in the intended sense here are the necessary ingredients for deliberation.

Our ordinary ascriptions of reasons are context-dependent in more than one way. First, the term ‘reason’ can be used in different senses in different contexts. As earlier noted, for example, what we mean by ‘reason’ is subjective reasons in some contexts but objective reasons in others, which is explained by context-sensitivity of the assignment of the agent’s \((R, S)\). Similar stories can be told of other familiar distinctions between reasons like that between motivating and normative reasons.

Context also plays a role in determining which item in a piece of reasoning is pragmatically salient. When we explain or evaluate an action by citing a reason, we don’t usually give every detail of the agent’s reason or reasoning. Context tells us what we are assuming and hence what is not at issue. In most cases, it is enough only to mention some item in the agent’s reasoning that is most significant in the given context. So usually we only cite a belief or desire of the agent or some noteworthy feature of the external circumstances. Loosely speaking, then, any premise of an instance of reasoning may be said to state a reason. In a word, a reason is a premise of reasoning. More precisely, however, the set of all the premises in the given instance of reasoning comprise a reason. This is reflected in the way we talk about reasons: We often say that our beliefs, desires, or some facts *give* or *provide* us with a reason to act in a certain way, rather than that they *are* themselves a
reason. They constitute a reason, on the present approach, along with other items which are not conversationally salient in the given context. We sometimes say that the fact (or belief) that Billy will be at the party is a reason for Mary to go to the party, and sometimes that Mary’s desire to see Billy is a reason for her to go to the party. However, they are just two different ways of talking about the same reason.

I now turn to what the present approach can say about reason’s normative and explanatory dimensions. As we noted, a person may have multiple reasons to act in a certain way. Consider again Mary who has more than one reason to go to the party. Her reasons can be represented thus: \( r_1, s_1, \phi \), where \( r_1 \) represents the belief that there will be gin at the party, and \( s_1 \) the desire for gin, and \( r_2, s_2, \phi \) where \( r_2 \) represents the belief that Billy will be at the party, and \( s_2 \) the desire to see Billy. Mary also has a reason rather to stay home, say, \( r_3, s_3, \psi \), to finish her paper, but let’s suppose that she totally forgets about her paper and cares only about gin and Billy. So the attitudes \( r_1, s_1, r_2, s_2 \) are all in Mary’s actual \( \langle B, D \rangle \) and constitute her motivating reasons to go to the party. Here there is a possibility, as Davidson (1963) stresses, that she goes to the party only for one of those two reasons she has. Let us call a reason that actually explains why the agent did what she did an explanatory reason. We may well, then, ask what is the thing that is present in an explanatory reason but absent from merely motivating reasons that makes the former explanatorily efficacious. Causal theorists would explain the difference of their explanatory status in causal terms, for instance, by saying that \( r_1 \) and \( s_1 \), but not \( r_2 \) and \( s_2 \), caused Mary to \( \phi \). This seems to be a natural way to accommodate causal approaches within my framework.

What I want to stress is rather this. On the present picture, reasons are essentially or inherently normative, in that the premises in a given instance of reasoning must stand in appropriate normative relation to the action in the conclusion of the reasoning. So, every reason is a normative reason. Now, as we have seen, some reasons turn out to be explanatory reasons, while others do not. But this need not, and should not, be taken to mean that normative and explanatory reasons are of two different kinds or in ontologically different categories. Explanatory reasons are just reasons that satisfy some further conditions—for
example, reasons whose elements caused the agent to act as she did. Thus, every reason is both normative and also the sort of thing that can be an explanatory reason. There is no need of bifurcating the concept of a reason.

My account of reasons also helps us to get clear on two distinct normative dimensions involved in reason claims. Suppose you see Billy set fire to a cat. He wanted to know whether the cat’s hair is burnable, or maybe he just hates the cat. Then, there was a reason for Billy to set fire the cat; his action is an action done for a reason after all. Now, we might also well ask: Is there really a reason for him to burn the cat? The thought here would be that neither the fact that Billy hates the cat nor his desire to learn about the cat’s hair really counts in favor of his burning the cat. Indeed, there seems nothing that can really be said to be a reason to burn the cat in a fully normative sense. However, note that there is a sense, though a somewhat attenuated one, in which the fact that Billy hates the cat, or Billy’s desire to learn about the cat’s hair, does count in favor of his burning the cat. That’s why we find it natural to say that such a fact or attitude could be Billy’s reason for what he did. We wouldn’t say a similar thing about the fact that, say, Billy hates dogs or his belief that cats are mammals—none of them favors Billy’s burning the cat even in an attenuated or permissive sense. And it is this thin sense of reason that I propose to capture in terms of what I take to be the minimal, basic sense of reasoning.

One way of describing the situation will be this: There really were reasons for Billy to burn the cat, but there was no good or justified reason to do so. If we speak of ‘reasons’ in this way, we allow that there could be bad or silly or unjustified reasons as well. By contrast, if we use ‘reason’ in a fully normative sense, a ‘bad reason’ is like an oxymoron. Being good is already built in the idea of a reason, and so phrases like ‘good reason’ would be pleonastic. Something’s being a reason in a fully normative sense, thus, means more, and so takes more to be true, than its being a reason in the thin sense. My strategy has been to first give a rough analysis of reason claims in the thin sense and then make it into a general schema to cover reason claims in a fully normative sense as well.
For a putative reason ⟨r, s, φ⟩, then, there are two questions we can ask concerning its normativity. First, we can ask whether it really embodies appropriate normative relation required for a reason. That is, we can ask whether r and s really are R-related to φ. If they are, the putative reason counts as a reason—so we may regard the question as a constitutive normative question about reasons. Once it turns out to be a reason, we can also ask whether it is a good or justified reason. We can ask whether it is a reason in the so-called standard normative sense, or, more generally, we can ask a number of evaluative questions about reasons. Apparently, a reason could be good or bad in a variety of ways, and I believe much of our normative discourse in general can be understood in terms of claims about reasons, particularly claims about in what ways the reasons are good or bad.

6. Some Metaphysical Issues

In her recent paper entitled “Acting for a Reason,” Christine Korsgaard characterizes the “question about the ontology of reasons” as a question of “what a practical reason is: that is, what we are referring to when we talk about ‘the reason for an action’” (2009, 207). Or, “what sorts of items count as reasons for action” (2009, 208)? On the view presented here, reasons do not form a special or sui generis ontological kind. Reason claims are claims about reasoning in a rather ordinary sense. When we deliberate, explain or evaluate action in terms of reasons, we are just dealing with pieces of reasoning. Of course we can, and should, ask what reasoning is and what kinds of things the items in reasoning are. But these questions are not new, nor particular for moral theorists. At the very least, reasons are not metaphysically mysterious creatures—if or to the extent you feel comfortable about reasoning, you should feel comfortable about reasons.

So I agree with Scanlon when he says, “The question ‘What is a reason?’ is misleading insofar as it suggests that reasons are a special ontological class” (1998, 56). Now, he goes on to say, “What is special about reasons is not the ontological category of things that can be reasons, but rather the status of being a reason.” By ‘the status of being a reason’ Scanlon means the relation of being a reason for something, namely the relation of “counting in favor of” something. His point is that distinctive metaphysical concerns about reasons, if
there are any, arise from the normative relation of being a reason for or counting in favor of (rather than from things that are reasons). For example, it might be asked whether truths about reasons commit us to normative relational properties that would be metaphysically odd.

I agree with Scanlon’s point, word for word. But a primary metaphysical concern about reasons for my view is different from Scanlon’s, because what I would mean by ‘count in favor of’ is different from his meaning. As I mentioned at the start, Scanlon believes that the notion of a reason in the standard normative sense is not informatively analyzable. He also believes that other normative relations or properties can be built from the relation of being a reason for, or counting in favor of, which he takes to be explanatorily basic—hence, “Reasons Fundamentalism.” The question, then, is exactly how we should view the basic normative relation of counting in favor of. In parallel fashion, on my proposed approach, claims about reasons are built from claims about reasoning, and we have taken as given the appropriate normative relation that underwrites reasoning, namely the $R$-relation. The question then remains exactly how we should view the $R$-relation.

But the parallel is superficial. As noted above, being a reason for in the standard normative sense means more than being a reason for in the thin sense, which I analyze in terms of reasoning (or the $R$-relation). While Scanlon takes the former as primitive, I provide some explanatory story about how it might be understood in terms of the latter. By taking reasons as basic, Reasons Fundamentalists make them unexplainable, and leaves unexplained many things about reasons that need explaining. One feature that people often find attractive about Reasons Fundamentalism is that it preserves the irreducibility of the normativity of reasons. As it stands, my approach leaves the normativity of reasons unreduced. But by locating the basic normative element inside a process of reasoning, it gives us a way of talking about reasons that allows us to explain some key features of reasons that Reasons Fundamentalists can only treat as brute.
CHAPTER 2

Action Explanation, Rationality, and Mental Simulation

In this chapter, I sketch a simulation-based account of action explanation. My approach is a normative one in that it recognizes an important role for normativity in analyzing the explanatory force of action explanation. But it is also ‘Davidsonian’ in that it can be seen as a simulationist reconstruction of Donald Davidson’s views on action explanation and rationality. On this account, briefly, reasons are rational (or rationalizing) causes in which ‘rational’ is to be understood in terms of being simulatable. I defend two main claims about this notion of rationality: First, rationality is the constitutive normativity of reasons, the normativity in virtue of which an attitude counts as a reason for an action. Second, rationality sets limits on the actions or attitudes that we can understand, such that we can only understand the actions of those agents who share the same principles of rationality with us.

1. Introduction

We explain our actions or attitudes by citing our psychological states such as beliefs, desires, hopes, and the like. Why did Billy open the fridge door? Because he believed there was a beer in the fridge, or, perhaps just simply, because he wanted a beer. We deal with folk psychological explanations of this sort every day. They are fundamental to the way we understand ourselves and our fellow human beings.

A distinctive feature of this kind of explanation is that it explains an action by revealing the agent’s reason for performing the action. When you explain Billy’s action by citing his mental states, or reasons, it shows the action’s point or rationale. In this regard, folk psychological action explanation differs from the sorts found in sciences like physics and biology. It also differs from other sorts of commonsense explanations of events, for example, the sort you might use to explain why the plane crashed last night—obviously, a plane
crash has no point or rationale. Then, exactly how do such action explanations work? What is it about a reason that makes it explanatory of the action to be explained? What makes a given reason-giving explanation (or just reason explanation) an adequate explanation, and what is the source of explanatory understanding when such explanation is provided? These are the questions to be discussed in this chapter.

Around the middle of the last century, Carl Hempel and William Dray debated this very issue. The debate was carried on in terms of explanations of historical events, but it was actually a dispute about the nature of explanation of human actions in general. In a broader context, the debate was also about scientific methodology. Hempel claimed that explanations in the human and social sciences, which deal with human agency and use the same conceptual framework of folk psychology, do not differ in their character from law-based explanations as found in the natural sciences. On Hempel’s view, folk psychological explanations are basically nomological, and the nature of understanding is basically the same in all areas of empirical scientific inquiry.

Some philosophers, especially in the Verstehen-hermeneutics tradition, objected to this view and took a fundamentally different approach to action explanation. They often distinguished between explanation (Erklären) and understanding (Verstehen) of phenomena—as the aims of the natural sciences and the human sciences respectively—and they held that understanding in the human sciences essentially differs from explanations in the natural sciences. R. G. Collingwood (1946), for example, argues that historical understanding requires not empirical laws but a distinctive first-personal methodology of “re-enactment.” Dray (1957) develops this idea of reenactment or “empathetic understanding” and proposes the “rational” model of action explanation. On this approach, there is a distinctively normative aspect to reason explanations, and to understand an action, we must identify ourselves sympathetically with the agent and reenact the agent’s calculation or rationale for the action.

All normative, empathy-based theories along this line, however, face a forceful objection from Hempel. Simply put, the objection is that normativity or empathetic reenactment
has nothing to do with the “logic” of action explanation. Hempel’s criticism has been generally considered so decisive that the idea of empathetic understanding never gained a secure toehold in philosophical debates, though it has been recently revived in a different context, by simulation theorists about folk psychology.

I believe that there is something importantly right about Collingwood’s and Dray’s basic insights. In this chapter, I sketch a version of an empathy theorist or simulationist account of reason explanation. The account presented here is a kind of normative approach in that it does justice to the importance of normativity in analyzing the explanatory force of action explanation. But it is also ‘Davidsonian’ in that it can be seen as a simulationist reconstruction of Donald Davidson’s views on action explanation and rationality. On my proposed account, briefly, reasons are rational (or rationalizing) causes in which ‘rational’ is to be understood in terms of being simulatable. So the present approach can be viewed as an attempt to put causality, normativity, and empathy (or simulation) in their proper places in an analysis of action explanation. These components will be addressed in sections 3, 5, and 6 respectively. I will begin with a brief look at the old debate between Hempel and Dray. The whole discussion in the chapter can be thought to concern what Dray has to say in response to Hempel.

2. The Dray-Hempel Debate

As is well known, Hempel developed the “covering-law” model of scientific explanation. On this model, to explain an event is to give a law-based derivation of the event and thereby show that the event was to be expected under the circumstances. Hempel maintains that this holds for all explanations of individual events in any domain, including intentional human actions. When we explain an action by citing the agent’s beliefs and desires, for example, Hempel (1942, 1965) holds that we are appealing to some empirical psychological laws stating how agents with such beliefs and desires will generally behave. On this view, therefore, an agent’s psychological states explain his action by making it expectable or predictable on the basis of laws.

The picture Dray presents on the issue is radically different. Dray writes:
The goal of such explanation is to show that what was done was the thing to have done for the reasons given, rather than merely the thing that is done on such occasions, perhaps in accordance with certain laws (loose or otherwise). The phrase ‘thing to have done’ betrays a crucially important feature of explanations in terms of agent calculations…For the infinitive ‘to do’ here functions as a value term. I wish to claim therefore that there is an element of appraisal of what was done in such explanations; that what we want to know when we ask to have the action explained is in what way it was appropriate. (1957, 124; emphasis original)

On this approach, reason explanation makes an action intelligible by showing that “what was done was the appropriate or right thing to have done” (1957, 126). In short, normative appraisal, not nomic expectability, is the source of explanatory understanding in action explanation.

Dray also assigns an essential role to empathy or empathetic reenactment in action explanation. According to Dray, to understand why an agent did what she did, we need to put ourselves in the agent’s position, or “empathetically identify” ourselves with the agent, and “re-enact” the agent’s thought or deliberation in our own mind. To understand why Billy opened the fridge door, for example, we project ourselves imaginatively into his situation and reenact his thought in our mind, asking ourselves what is the thing to do in his situation. Reason explanation, Dray says, requires a “reconstruction of the agent’s calculation of means to be adopted toward his chosen end in the light of the circumstances in which he found himself. To explain the action we need to know what considerations convinced him that he should act as he did” (1957, 122).

Questions arise, however, as to how normative appraisal and empathetic reenactment can be relevant to the “logic” of action explanation, to use Hempel’s phrase. First, as Hempel admits, it does seem that reason explanations have normative or evaluative overtones. But plainly, evaluation or justification is one thing and explanation is another. How then can showing why an action should have been done explain why it was actually done?
In discussing Dray’s account, Hempel writes: “the information that agent A was in a situation of kind C and that in such a situation the rational thing to do is x, affords grounds for believing that it would have been rational for A to do x, but no grounds for believing that A did in fact do x... the appraising function which Dray considers essential for rational explanation has no explanatory import” (1965, 471; emphasis original).

Second, Hempel also admits that the method of empathy may be useful, and is perhaps what we frequently do, in explaining others. But he argues that empathy does not in itself constitute or generate an explanation. Rather, Hempel claims, it is “essentially a heuristic device” for suggesting certain psychological hypotheses which might serve as explanatory principles (1965, 239). The point is that empathy does not guarantee the correctness of an explanation to which it leads, and it is irrelevant to the explanatory force of action explanation.

Dray denies that empathy is just a heuristic device. He writes: “The point of the ‘projection’ metaphor is... more plausibly interpreted as a logical one. Its function is not to remind us of how we come to know certain facts, but to formulate, however tentatively, certain conditions which must be satisfied before a historian is prepared to say: ‘Now I have the explanation’” (1957, 128; emphasis original). However, Dray failed to make his case, leaving many important questions unanswered.

To sum up, for Hempel, to give an action explanation is to show that the action had to happen as it did, which requires its subsumption under laws. For Dray, to give an action explanation is to show that the action was appropriate in light of the agent’s reasons, which requires reenactive empathy. Hempel’s criticism of Dray has two prongs, one about normativity and the other about empathy. Hempel rejects both the idea that action explanation consists in showing that the action was appropriate, and the idea that that action explanation requires empathy.

Here is, then, my project on behalf of the empathy theorists, which falls into two steps: The first is to identify the normative dimension of action explanation that is relevant to its explanatory force; the second is to show why empathy is needed in accounting for this
normative dimension. One quick point is in order. Although Dray lays stress both on the role of empathy and the normative aspect in reason explanation, he never makes it clear exactly how the normative aspect is related to empathy. Note that even if it is agreed that normative appraisal is necessary for reason explanation, one could insist that empathy is at best a useful heuristic in finding out whether the action was appropriate. Conversely, even if action understanding requires empathy, it does not seem to follow that the understanding involves or requires normative evaluation. What then is the supposed connection, if there is any, between the two supposed elements of action explanation? I believe that only by understanding the right connection between them can one plausibly defend both ideas. Establishing it, as we will see, also gives us a way of meeting Hempel’s objection.

3. Reasons as a Species of Cause: Making Dray Davidsonian

We can put our question about the nature of action explanation this way. When we explain an action by citing the agent’s psychological state, what is it that makes his psychological state an explanatory reason for the action? The first point of Hempel’s criticism of Dray was that the justifying or evaluative role of reasons plays no role in answering this question. Crudely put, the point is that showing that an action is the appropriate thing to do does not entail, hence does not explain, that the agent actually performs that action. It might be thought that Hempel here presupposes that nomic entailment is the only explanatory relation for all explanations, simply rejecting Dray’s idea that normativity can generate explanatory understanding (e.g. Kim 2010, 136). This reading, however, does not fully appreciate Hempel’s basic point, which I take to be that Dray’s rational explanation simply does not answer the ‘why’ question. Showing that some of the agent’s beliefs or desires make it rational or appropriate for him to act in a certain way is perfectly consistent with his not acting that way. Certainly we often fail to do the appropriate thing. How, then, can merely showing that the agent’s action was the thing to do in light of his mental states explain why he in fact acted that way?

But recall that there is more to Dray’s rational explanation than just showing that the action to be explained was appropriate. Dray also emphasizes reconstruction of “the agent’s
calculation” by using the method of empathy. It might be argued, then, that in order to qualify as a rational explanation, the explanation must somehow represent the agent’s calculation *correctly*, just as in order to qualify as a Hempelian D-N explanation the explanans must be *true*. On this construal, an action explanation counts as a true rational explanation only if it correctly represents the agent’s normative deliberation.

I don’t think, however, this kind of move works. It is possible that an agent acts in accord with the result of his deliberations without acting on the reasons that figured in the deliberations. Imagine, for example, a man who deliberates and judges that he should rescue a drowning child, and dives into the river, but he does it in fact only out of a desire to impress his girlfriend. In general, we can coherently suppose that an agent may have a number of reasons, any of which makes his action the thing to do, but performs the action for only one of those reasons. In such a case, then, how can showing that the action was the thing to do in light of the agent’s calculation show that he acted as he did *because of* the reason that corresponds to that calculation?

As is now familiar, these are cases of the sort Davidson uses to argue for a causal conception of action explanation. Everyone agrees that crucial to the connection between a reason and an action it explains is the condition that the agent performed the action *because* he had the reason. Any adequate account of action explanation should be able to explain the force of this ‘because.’ And any plausible story about the force of the ‘because’ should be able to distinguish a reason *for which* an agent acts from a reason he merely has for his action. Now, what other than causation could distinguish between them and capture the force of the ‘because’? This is the so-called Davidson’s challenge to non-causal theorists. Davidson argues that causalism, as we might call it, alone promises to give an account of “the ‘mysterious connection’ between reasons and actions” (1963, 11).

One way of putting Hempel’s objection is that Dray’s model does not explain why the action was in fact performed rather than not performed. If this is right, *a fortiori*, it does not explain why the action was performed for one reason rather than for another. How should Dray counter this line of objection? The empathy approach is generally considered
an anti-causalist position, and it is true that all empathy theorists have actually been anti-causalists. The reason seems to be clear. The causalist holds that a reason explains an action only if it is a cause of the action, so that reason explanation turns out to be “a species of causal explanation” (Davidson 1963). Once it is granted that reason explanation is a type of causal explanation, however, it is not at all clear what role empathy can play in a “logical” analysis of action explanation. That also undermines the idea that reason explanations are essentially different from causal explanations in the natural sciences. It may seem then that empathy theorists had better be anti-causalists.

I believe, however, that the basic ideas of the empathy approach are consistent with the causalist thesis that reasons are causes. As I take it here, the causalist thesis is a metaphysical thesis about the connection between reasons and the actions they explain. Note that the thesis as a metaphysical one is enough to meet Davidson’s challenge. It gives us a way to distinguish explanatory reasons from mere reasons the agent has—explanatory reasons are reasons that cause the actions. But a metaphysical claim about how reasons are connected with actions does not give the whole story about how reasons explain actions. In particular, that a reason explains an action and that the reason is a cause of the action do not entail that the reason explains the action in virtue of being its cause. I do believe that a reason explains an action only if it is a cause. But I don’t think that the explanatory force of a reason is exhausted by its causal efficacy; it partly comes from elsewhere. On this line of thought, reasons are indeed a species of cause, but this does not mean that reason explanations are simply causal. Reason explanations, though causal, are irreducibly different from the kind of causal explanation in the natural sciences.¹

If empathy theorists can accept causalism, I think they should endorse it. If they don’t, I don’t think they can meet Davidson’s challenge. Indeed, I will shortly consider extant empathy theorist, or simulationist, but non-causalist accounts of action explanation, and argue that they don’t meet Davidson’s challenge. All this, however, looks more like a

¹I also believe that this is a close description of Davidson’s view. Davidson holds that there is an irreducible difference between folk psychological explanations that involve propositional attitudes and explanations in sciences like physics that involve strict laws.
victory for causalists than for empathy theorists. Once it is admitted that a reason explains an action even partly by causing it, an urgent task for empathy theorists is to show why causation is not all there is to reason explanation, and why we still need empathy. The more interesting and important issue is whether causalists need empathy, rather than whether empathy theorists need causation.

Note also that simply introducing the causal element into the empathy approach gives no answer to Hempel’s criticism that neither normativity nor empathy is relevant to the logic of action explanation. Rather, it reinforces the criticism. The central task for the empathy approach, then, is to show why normativity and empathy are needed in an analysis of reason explanation. Before we move on, though, let us briefly pause and consider the leading extant empathy-based accounts of reason explanation proposed by Robert Gordon and Jaegwon Kim in recent years.

4. Gordon and Kim on Action Explanation and Simulation

The basic idea of empathetic understanding has made a strong comeback through the work done by the simulation theorists in philosophy of mind and psychology. Narrowly conceived, the simulation theory is a theory about mind reading. More broadly, however, it is a theory about our everyday, folk psychological competence in general. Probably it is best understood by contrast with its rival approach, i.e., the theory theory of folk psychology. According to the theory theory, we can think of folk psychology as a kind of empirical theory about the way people think and behave. Psychological terms like belief and desire, then, can be regarded as “theoretical terms embedded in a folk theory which provides an explanation of people’s behavior” (Stich 1983, 19). On this approach, our engagement in the ordinary practices of interpreting, predicting, or explaining others’ thoughts and behavior draws on a rich body of tacit theory.

The simulation approach challenges this view. It denies the idea that our understanding of others proceeds primarily by using an empirical theory, or body of information, whether tacit or not. According to the simulation theory, we use the resources of our own minds to simulate the mental life of others. In predicting or explaining others’ actions and thoughts,
for example, we imaginatively put ourselves in others’ shoes, run our own decision-making system hypothetically, and ascribe the outcome to the other. On this approach, we are simulators rather than theorizers. It is our ability to engage in mental simulation in imagination that is essential to our folk psychological understanding of others.2

The simulation theory is standardly considered as a descriptive account of our folk psychological practice. Its claims about action explanation, for example, mainly concern how our psychological explanations are actually made and what cognitive mechanism underlies our explanatory practice.3 It is not obvious, then, how the simulation theory thus understood can be pertinent to our concern with the logic of action explanation. It might well be part of a true description of how the folk arrive at everyday explanations, but the question on the table is rather what makes such explanations correct.

It might be thought that simulation can provide more than a description of our explanatory practice and should be part of a story about the logic of action explanation. This is Gordon’s position. Gordon (2000, 2001) tries to show how simulation can yield an account of what being a correct action explanation consists in. In his recent paper, Kim (2010) also provides a simulation-based account of action explanation, without committing himself to one side or the other in the simulation vs. theory debate. Kim attempts to give a plausible reconstruction of Dray’s model and tries to show “how normativity can generate explanatory understanding—or, how the normative relation of justification can be an explanatory relation” (2010, 147). In this section, I consider first Gordon’s account and then Kim’s, and argue that neither has resources to meet Davidson’s challenge.

In giving his own simulationist account, Gordon acknowledges that failure to answer Davidson’s challenge is a major shortcoming of its predecessor views like Dray’s. So his announced aim is to show how action explanations conforming to his account can “perform an explanatory function beyond merely portraying the action as reasonable from the agent’s

---

2 This approach was introduced independently by Gordon (1986) and Heal (1986) and soon joined by Goldman (1989), though see also Dennett (1981) and Stich (1981). Heal (1986) uses the term ‘replication’ instead of ‘simulation.’

3 For example, focusing on psychological ascription rather than explanation, Alvin Goldman makes it explicit that his “descriptive epistemology” concerns “naive processes that produce mental-state beliefs, whether or not these beliefs are justified, rational, or specimens of knowledge” (2006, 10).
point of view.” He goes on to say, “They surely do answer the ‘Why?’ question, and they may have the form ‘A X’d because R,’ where R states a reason of A’s for X’ing” (2000, 63).

Now, Gordon’s account can also be construed as a version of a counterfactual account. He puts his idea of how a reason explains an action this way (2001, 189):

Reason R explains A’s X’ing in that, without R, A would not have X’d.

Consider Gordon’s example. When I braked for a red light, I had another reason to brake; I had been going ten miles over the speed limit. But I braked because of the red light, not because of my speeding. It is the red light, or my belief about the red light, that explains why I braked. What is it, then, about this belief that makes it explain my braking? What accounts for the difference in the explanatory status between the two reasons for my braking? On Gordon’s view, it is explained by an asymmetry between the following counterfactuals:

(C1) If there hadn’t been a red light, I would not have braked;
(C2) If I hadn’t been going ten miles over the limit, I would not have braked.  

Here (C1) is true, while (C2) is false, and that’s why my belief about the red light, not my belief about speeding, is explanatorily efficacious in explaining what I did. To this, causal theorists will say that the asymmetry between (C1) and (C2) just reflects the difference in the causal status between the two reasons I have. That is, it is my belief about the red light, not my belief about speeding, that caused my braking, which accounts for the difference in their explanatory status.

4Here the antecedents of the counterfactuals are not couched in terms of the agent’s mental states; Gordon identifies reasons as a certain kind of consideration rather than as mental states. This does not affect my discussion to follow.
Gordon rejects this causalist view. He claims that there are two ways of evaluating counterfactuals, causal and rational, and that rational evaluation, which requires a simulative procedure, is what we need in an account of action explanation. Gordon (2001, 177) describes “a simulation procedure for determining what a person would have done under counterfactual conditions” as follows: “Wondering what agent A would have done at time \( t \) if at \( t \) it had been the case that \( p \), I first simulate A at \( t \). Then, still simulating A, I pretend that \( p \), in effect simulating A in the world thus transformed; then, in the role of A in the counterfactual condition \( p \), I decide what to do” (2001, 186). Gordon’s idea seems to be this. To see whether a reason, R, explains the agent’s action, X, we need to evaluate some relevant counterfactuals, that is, what the agent would have done if he had not had the reason R. And for this, we need a simulation: We put ourselves in the agent’s place, where things are as similar as possible except that the agent does not have R, and then decide what to do.

I find this idea very puzzling. Briefly put, Gordon’s simulation procedure consists in seeing that if I were you, and I hadn’t had R, I would not have done X either. Then how can this sort of procedure “determine” what the agent would have done under the counterfactual circumstances? Does this mean that what the agent would have done under certain counterfactual situations depends on my simulation of him? This seems highly implausible. How can my simulation of the agent’s mental states be relevant to the conditions for them to be his reason for the action? It is worth noting that Gordon’s example is the first-person case, from which I think his account gains some spurious plausibility. Gordon might then reply that his simulation procedure is to be understood as a procedure to determine what I would have done under some counterfactual circumstances. So his account, Gordon might claim, is an account of our self-explanation of our own action, namely, an account of what makes a certain mental state of mine my reason for the action.

However, it still seems implausible to hold that what I would have done under some counterfactual situation is determined by my ex post facto simulation at a later time. It seems to me that Gordon’s simulative procedure is plausible as an empirical hypothesis.
about our actual practice of predicting (rather than explaining) our own actions in hypothetical situations. Indeed, Gordon writes, “the distinction between the causally interpreted conditionals and the rationally interpreted conditionals is not to be understood in terms of a theoretical difference. It is a matter of a procedural difference, a difference in what we do in interpreting the corresponding conditionals” (2001, 185; emphasis original). Here Gordon seems to mean that his simulative procedure is to be understood not as an account of the truth conditions of some sort of counterfactuals, but as an account of what kind of process we are engaged in when we perform counterfactual reasoning about what we, or others, would have done under some hypothetical circumstances.

If so, however, does Gordon have an answer to Davidson’s challenge? On his account, that a reason explains an action consists in the fact that without the reason, the agent would not have done the action. Then the question we should ask is what makes such a counterfactual true, which would account for the explanatory force of the reason, rather than what is actually going on in our mind in evaluating it. Gordon has the following observation: “The presence of the red light was sufficient (in the context) to move me to step on the brake; my going ten miles over the speed limit was not” (2001, 182). He also writes, “I rejected the account that posits a causal chain... But how are these counterfactuals to be understood, if not by such a chain? As the phrase ‘sufficient to move me’ suggests, we seem to be talking about a sufficient reason or rational conditions. That there was a red light just ahead was, together with contextual facts, reason enough to move me, whereas my speeding was not” (2001, 184; emphasis original). Gordon seems to claim that what makes (C1) true and (C2) false is the fact that my belief about the red light was “rationally sufficient” to move me, while my belief about speeding was not. However, shouldn’t we have considered a case in which my speeding was also rationally sufficient to move me in order to address Davidson’s challenge? What if an agent has more than one rationally sufficient reason but she acts on just one of them? In such cases, Gordon’s account seems to have nothing to say about what makes a reason an agent has a reason for which she acts, a reason that explains her action. I conclude that Gordon’s account is better viewed as an
account of how we explain actions, rather than how reasons explain actions. I now turn to Kim’s account.

Kim (2010) defends the empathy theorist ideas and offers a “normative” and “agent-centered” model of action explanation that takes the agent’s first-person perspective as central.5 One of Kim’s main motivations is that causal-nomological approaches do not give an adequate account of how our self-understanding works. His starting point is the observation that the way I understand my own action is “to know, or be able to reconstruct, the practical deliberation that led me to decide to do something (and actually do it),” not subsuming it under an empirical psychological generalization (2010, 139). He then argues that similarly, a third-party’s understanding of an agent’s action consists in knowing or recapitulating the agent’s deliberation leading to the action. Kim’s argument can be summed up as follows:

(P1) The way I understand my own actions is to know or reconstruct the practical deliberation that led me to do what I did.

(P2) The way I understand my own actions is pretty much like the way you understand your own actions.

(P3) The way I understand your actions is pretty much like the way you understand your own actions.

(C) The way I understand your actions is to know or reconstruct the practical deliberation that led you to do what you did.

For example, Kim asks, “What must I, the third-person explainer, do to gain an understanding of why Mary did what she did?” Kim writes, “I simulate the agent’s practical deliberation by imagining myself in the agent’s situation and asking myself what is the appropriate thing to do—appropriate for me to do,” and that “The simulatator’s understanding of another’s action lies precisely in the fact that his deliberation, using his deliberative principles, reaches the same conclusion as the agent’s deliberation based on her action principles. That is how he can say ‘Now I see why she did it; I would have done the same in her situation!'” (2010, 144-145).

5For the importance of the first-person perspective in understanding actions, see also Kim (1984, 1998).
This is a cogent line of thought. Kim does justice to the importance of self-understanding in explaining actions, and I believe there is something right about Kim’s account. However, the question I have concerns, again, the scope and status of his account. We all should agree that if we come to know the deliberation that led the agent to do what she did, we understand her action. But I think we should ask: What is the source of the understanding or explanatory import when we understand an agent’s action this way? What is it about knowing the agent’s deliberation that gives us an understanding of her action?

More specifically, here is how Hempel’s objection might be pressed. Let us grant the importance of “knowing the deliberation that led the agent to do what she did” in understanding actions. What is still unclear is why we need a simulation. Can’t we know the agent’s deliberation without any kind of simulation procedure? More importantly, it seems we can perfectly well explain or understand actions which we think we would never have done—I understand much of what, say, Hitler did. Why then does knowing the agent’s deliberation require a simulation, especially one to the effect that I would have done the same? How can seeing that I would have done the same in his situation be relevant to the explanation of why he in fact did what he did?

This line can be pushed further, drawing on Davidson. It seems that when we know the agent’s deliberation that led her to her action and thereby understand her action, much of the explanatory import comes from causal information. Notice that ‘led’ is a pretty overtly causal term. Note also that knowing the agent’s deliberation generates an explanation only if the deliberation is the deliberation that actually led him to his action. Once we note this, “knowing the agent’s deliberation that led him to do what he did” sounds almost like “knowing the agent’s reason that caused him to do what he did.”

Indeed, unless the ‘led’ is understood as causal, I don’t think that Kim’s account has resources to distinguish a reason for which an agent acts from a reason she merely has. It is noteworthy that Davidson’s challenge can be equally raised for self-understanding. Suppose I somehow forgot why I took my wife’s car instead of mine. I try to figure out why I did that and so simulate myself of this morning. I start with the belief that the gas

\[6\] The example is from Kim (2010, 131).
price is rocketing and that my wife’s car gets better mileage, which I actually believed in that moment, and arrive at the decision to take my wife’s car. I feel I now see why I took my wife’s car this morning. But it may turn out that I was just self-deceived in my simulation; the reason I took my wife’s car was actually because I saw my car had a flat tire. More importantly, even if we suppose my simulation was in fact veridical or successful—my belief that my wife’s car gets better mileage was really my reason for taking her car—the fact that that belief explains my action seems to have little to do with how my simulation actually went. What does the explanatory work seems to be the fact that that belief really caused me to take my wife’s car.

It may be true that simulation is what we must do to understand or explain an action. Recall the way Kim poses his question: “What must I, the third-person explainer, do to gain an understanding of why Mary did what she did?” Indeed, I believe that simulation should be part of a plausible “epistemology of rationalizations” (Kim 2010, 134)—notice also that Kim’s key term is ‘understanding’ rather than ‘explanation.’ However, this does not put to rest Hempel’s criticism, to the effect that although there is something right about the simulation approach, “the element of truth in such an account is not a point of logic; it is a mixture of psychological description and methodological precept.”

5. The Constitutive Normativity of Reasons: Rationality

In light of the discussion so far, it might seem that neither normativity nor empathy is necessary for an account of reason explanation. Indeed, I believe Hempel is right that justification (in its usual sense) is not relevant to the explanatory force of reasons, and that simulation alone, even a successful one, provides no explanation. So I believe Davidson is right that a reason explains an action by causing it. But this does not mean that reason explanations are simply causal, nor that empathy or simulation has no place in an analysis of reason explanation. I believe that there is an inherently normative aspect in action explanation that is explanatorily relevant. One way of putting the idea is this: It is true that what makes a reason an explanatory reason is the fact that it causes the action, but causation is

---

This way of putting Hempel’s point is from Dray (1957, 12).
not all there is to what makes a *psychological state* an explanatory reason—to make it a reason in the first place, some sort of normativity is needed. In this section, I will try to identify this normative element. I then try to capture it in terms of simulation in the next section.

The kind of normativity I have in mind here is not what is involved in making an action *right* or *justified*, but what is involved in making a psychological state a *reason* for an action. For something to explain an action, it must be a reason for the action. Yet, on one reading, this is just obviously false. Consider Billy who tortures a cat because he hates her. You might say that the fact that one hates a cat is not at all a reason for him to torture her—that is never a reason in a fully normative sense of reason. But nonetheless his hating the cat nicely explains his action. In this sense, it’s not true that something explains an action only if it is a reason for the action. This is not the reading I intend, however. What I have in mind is the sense in which it is almost conceptually, or trivially, true to say that only reasons explain an action. They need not be a *good* reason, however, morally or rationally or whatever. Like Billy, we sometimes act for a truly bad reason, and bad reasons perfectly well explain our actions. It’s because bad *reasons* are still reasons. Thus, what it is about a reason (or psychological state) that is relevant to its explanatory status has nothing to do with what makes it good or bad reason, which may well be relevant to evaluating or justifying actions.

So there is a sense in which Billy’s hating the cat is a reason for him to torture her. Let us focus on this somewhat thin, permissive sense of reason. Now it is clear that not every mental state an agent has can be cited as a reason for an action in this sense. Mary’s belief that it is raining might be cited as a reason for her taking an umbrella. Her belief that she can protect herself from lightning with an umbrella might also be cited as a reason for taking the umbrella. Even her desire to give her son the rod with it might be cited as a reason for taking it. However, nothing like this would be said for her belief that, say, snow is white. The point is that for a mental state to be a reason for an action, it must be appropriately related to the action. The action must be in some way *desirable* or *attractive*...
in light of the mental state. Davidson expresses the idea this way:

[B]eliefs and desires tell us an agent’s reasons for acting only if those attitudes are appropriately related to the action… To serve as reasons for an action, beliefs and desires need not be reasonable, but a normative element nevertheless enters, since the action must be reasonable in the light of the beliefs and desires (naturally it may not be reasonable in the light of further considerations). (1978, 84)

Davidson says that an attitude is a reason for an action only if it makes the action reasonable. There is a sense, of course, in which Billy’s hating the cat never makes his action reasonable, though it is a reason. But nonetheless there is a sense in which it does make his action reasonable. In Davidson’s words, “there is a certain irreducible—though somewhat anaemic—sense in which every rationalization justifies” (1963, 9).

All this simply reflects how our reasons talk works in everyday explanatory contexts, none of which I take to be tendentious. As we have seen, the normative dimension of reasons I am tracing is much weaker than those involved when we usually evaluate or justify actions. We often do foolish things that are not appropriate at all. Even in such cases, however, we can perfectly well give reason explanations. In fact, an action can only be assessed as inappropriate or foolish when we can give a reason explanation—if we cannot, the action would be just unintelligible to us.8

The upshot is this: In order for an attitude to be a reason for an action, some appropriate normative relation must hold between them, in virtue of which the attitude counts as a reason for the action. The action must be in some way desirable in light of the attitude, or the attitude must render the action reasonable in some minimal sense. Let us use the term

8Consider Dray: “I do not wish to imply that anything that is explained on the rational model is thereby certified without qualification as the right, or proper, or intelligent thing to have done. In saying that the explanation must exhibit what was done as appropriate or justified it is always necessary to add the philosopher’s proviso: ‘in a sense’” (1957, 124; emphasis original). I take it that the “sense” Dray was trying to convey here is essentially the same as Davidson’s “anaemic” sense of justification.
“Rationalizing” relation to refer to such constitutive normative relations that hold between an attitude and an action for which it is a reason. In virtue of what, then, does an attitude serve as a reason for an action? Because it Rationalizes the action.

We should note that the Rationalizing relation must be explanatorily relevant if it is the constitutive normative relation that connects reasons with actions. As we have already noted, something explains an action only if it is a reason for the action. Now, something is a reason for an action only if it Rationalizes the action. Then, it follows, something explains an action only if it Rationalizes the action. If an attitude does not Rationalize an action, it is not a reason for the action and, a fortiori, is not an explanatory reason for the action.

Here, then, is how action explanation works. In order for an attitude to explain an action, the attitude must not only cause the action, but also Rationalize it. In a word, reasons are Rationalizing causes.9

On this view, what makes reasons explanatory is causation, period. That is, for a reason to explain an action, it is sufficient for the reason to cause the action; if it is a reason, it already Rationalizes the action. But this does not mean that causation is all there is to reason explanation. True, it is all there is to what makes a reason explanatory. But it is not all there is to what makes an attitude explanatory. For an attitude to explain an action, it is not sufficient for the attitude to cause the action. It must be a reason for the action. If it isn’t, it can’t be an explanatory reason for the action. Thus, Dray is right that there is an essentially normative component in action explanation that is explanatorily relevant. I believe Davidson expresses the same idea when he says, “If we were to drop the normative aspect from psychological explanations, they would no longer serve the purpose they do” (1991, 217).

9Compare Davidson: “Two ideas are built into the concept of acting on a reason (and hence, the concept of behavior generally): the idea of cause and the idea of rationality. A reason is a rational cause. One way rationality is built in is transparent: the cause must be a belief and a desire in the light of which the action is reasonable” (1974, 233).
6. Rationality as Simulatability: Making Davidson Simulationist

On the present view, if something is to be a reason for an action, it must Rationalize the action. What is it, then, for an attitude to Rationalize an action? When does an attitude Rationalize an action? Here is a simulationist approach: A set of psychological states Rationalizes an action if and only if we can simulate the process from those states to the action—that is, roughly put, when we imaginatively place ourselves in those states, we feel we are inclined to do that action. The resulting account of action explanation, then, would go like this: A set of attitudes explains an action only if (i) we can simulate the transition from it to the action; and (ii) it is a cause of the action.

Imagine, for instance, a situation in which we are told that Mary’s belief that snow is white caused her to carry an umbrella. We wouldn’t find it explanatory. It’s because we don’t find her belief about snow as a reason for her action—it does not Rationalize her action. Why doesn’t it Rationalize the action? Because, on my approach, we cannot simulate the transition from it to the action. There is a possibility, of course, that some of Mary’s attitudes are such that we can simulate the transition from those states together with her belief about snow to her taking an umbrella. Consider, for example, the following states: Mary believes it’s snowing outside; she also believes white things burn her skin; but she doesn’t want to burn her skin, and believes she can prevent burning by taking an umbrella. These states, together with her belief that snow is white, perfectly Rationalize her action; we can simulate the transition. Then, her belief about snow counts as a reason for taking an umbrella. And if it really is a cause of her action, it explains the action.

One thing that this example illustrates is that for any given instrumental belief/desire pair and action, we will have no difficulty in simulating the transition; when we project ourselves into such instrumental belief/desire state of mind, we feel we are inclined to do the action. This means that any instrumental belief/desire pair Rationalizes the relevant action and, therefore, counts as a reason for the action. Indeed, citing instrumental belief/desire pairs is the easiest and the prevailing way of explaining actions. It may be said then that instrumental belief/desire pairs deserve to be called ‘primary’ reasons (Davidson 1963).
This, however, does not commit us to the view that instrumental reasons are the only type of reasons that can explain actions. Consider the so-called “arational” actions that elude explanation by instrumental reasons. You are shouting and jumping up and down. Why you do that? You are so happy and excited. Why did you kick the door? You were just angry. Rosalind Hursthouse (1991) argues that cases like these do not fit Davidson’s belief/desire analysis of action explanation. According to Hursthouse, they are explained rather by an occurrent emotion, and if one tries to fit them into Davidson’s belief/desire formula, it over-intellectualizes them. The present simulationist approach provides a uniform treatment of such arational actions. Your excitement explains your shouting because it causes and Rationalizes your shouting; when we simulate you with such an input state of excitement, we feel we might shout as you did. In this way my approach can not only accommodate the centrality of belief/desire explanation in action explanations, but also account for the fact that we normally explain actions by citing various psychological states like emotion, mood, ignorance etc.

On this approach, an action is Rationalized or Rationalizable (henceforth ‘Rational’) when the agent has some mental states such that we can simulate the transition from those states to the action. So your shouting and jumping is Rational, and your kicking the door is a Rational action too. However, if we don’t find any such mental state in the agent, then the action would remain irrational and unintelligible to us. Imagine a student who knocks his head on the desk in the middle of the seminar. If, or to the extent that, we cannot see ourselves doing anything like what he did on any account of the matter, his action is not Rational and so we don’t understand the action.

The notion of Rationality involved here is far from what might be called perfect or ideal rationality. We don’t have any difficulty in understanding an action or attitude that is not fully rational. Indeed, nothing like rationality in this sense is constitutive of the notion of a reason. Consider, for instance, a woman who starves herself to death for the sake of a trim figure.\footnote{The example is from Gibbard (1990, 165).} Presumably, she has various other beliefs and desires including desire for survival. You might want to say that she is not rational, or that she doesn’t really have a reason to
starve. Surely, there is a sense in which she has no reason to starve, or that a desire for a trim figure is not at all a reason to starve, just as Billy has no reason to torture a cat. But this usage of ‘reason’—one which makes ‘good reason’ somewhat pleonastic—involves more than the thinner sense of reason I have in mind. More precisely, I think when one uses ‘reason’ that way, one means more by ‘reason’ than when we explain the girl’s action by saying that her reason to starve is her desire for a trim figure. Similarly, there is a sense in which her action is not at all rational. But the kind of rationality involved in such usage is not what I call the constitutive normativity of reasons, which concerns just what it is about her desire for a trim figure that makes it serve as a reason for her starving. The constitutive normativity of reasons, Rationality, has nothing obviously to do with what makes a reason good or moral or rational or whatever, nor with whether the action was done for the right reasons that requirements of rationality (in a thick sense of the term) impose. Rather, it is the normativity involved in what makes an attitude a reason for an action.\footnote{Given the common distinction between normative and motivating reasons (e.g. Darwall (1983), Dancy (2000)), one natural way of putting the point is this. The constitutive normativity of reasons is intended to capture the common denominator of normativity involved in both kinds of reasons, or normativity of something they have in common in virtue of which they both count as reasons.}

On the present approach, the starving girl’s action is Rational; we understand why she acted as she did. It is because we can simulate her—when I imaginatively identify myself with her, I feel I might act like her. Compare the starving girl with a man who has an intransitive preference. He prefers apple to banana, banana to orange, but not apple to orange. It seems that there is a significant difference between him and the starving girl. I would say, unlike the starving girl, we cannot simulate him. That is, he is irRational; we don’t understand him. If there were a woman who starves for a strong, intrinsic desire for self-destruction, she might well be irRational for the same reason—I don’t think I can simulate such an intrinsic strong desire for self-destruction.

Let us return briefly to the idea of simulation. As I use the term ‘simulation’ here, we can simulate a process or transition from certain states to an action (or to another state) just in case when we imaginatively put ourselves in those states, we feel we are inclined to do the action (or move to the other state). What does it mean that we “feel inclined to do”
something? I don’t think I can provide an analysis of this; I doubt that it’s possible. The way I prefer to think about it is that we should not hope to analyze it, but rather try to explain it through psychological theorizing. The rough idea is that there might be a significant kind of psychological state that corresponds to “feeling inclined to do” something. This state may have some similarity to the state of “intending to do” something, but will be weaker than that in some suitable sense.

7. Circles of Rational Intelligibility

On the approach sketched here, a set of attitudes explains an action only if it Rationalizes the action, and whether it Rationalizes the action depends on whether we can simulate the transition from the attitudes to the action. However, who are ‘we’ here? Doesn’t this relativize the notion of rationality involved, and hence of explanation, to individual simulators? Doesn’t the account look not so much like an account of the conditions required for an agent’s attitude to be an explanatory reason for her action, as an account of what is going on in our mind when we explain the agent’s action?

It should first be clear that the present approach is intended as an account of when an attitude explains an action and in virtue of what an attitude counts as a reason for an action. On this approach, it is true, whether a certain attitude an agent has is a reason for her action depends on whether we can simulate something in a certain way. (Thus, admittedly, when we say that the attitude explains her action or that it is her reason for the action, we are partly talking about ourselves!) But the condition about simulatability is about an objective relation that holds whether or not we simulate it, or however we actually simulate it. It is meant as part of the condition that must be satisfied for an attitude to be a reason for an action, not as a description about what is going on in us when we try to see whether it is a reason, nor as a heuristic in seeing whether it is a reason.

I am here adopting a strategy of the sort Gibbard (1990) adopts when he tries to explain what ‘accepting a norm’ is. According to Gibbard, to call something rational is to express a state of mind. Specifically, it's to express one’s acceptance of norms that permit that thing. Rather than define or analyze the ‘acceptance of norms,’ Gibbards suggests that there is a centrally important psychological state that roughly fits the notion of a person’s accepting a norm (1990, 55). His approach to rationality has a certain affinity with mine, though he takes his analysis to be an analysis of the ordinary sense of rationality.
Who is the ‘we’? It is indexical and refers to a population consisting of the simulator and those somehow like him. Here ‘like’ means like in what we can simulate. I would expect that if you are a thinking being, like me, then you and I are somehow very similar, in particular in simulating others; we are much the same kind of cognizer and agent. Suppose, however, that there is a certain pattern of attitudes and actions I can simulate but you can’t. If so, you are not like me; some of your actions might be totally unintelligible to me, and vice versa. So whoever the ‘we’ are, those who largely share the same patterns of simulation understand each other’s actions. Let us call such a population a “circle of Rationality.” Then the circle will place limits on which actions or attitudes we can understand.

Dray speaks of understanding the actions of “beings like ourselves.” In reconstructing Dray’s model, Kim suggests that this be taken to refer to agents who by and large share our “principles of action,” and calls the agents who share the same action principles the “circle of rational intelligibility.” Kim draws the conclusion that “we can only understand actions of agents like ourselves, in sharing the same action principle” (2010, 145). It will be illuminating to compare Kim’s circle of rational intelligibility with mine.

Let us first ask what Kim’s “principles of action” are. Kim speaks of “rules of deliberation and decision” or “rules and principles regulating practical deliberation.” On Kim’s account, for me to understand someone’s action, I need to simulate the agent’s practical deliberation by asking myself “what is the appropriate thing to do in the given situation” and “the outcome of my practical deliberation must match” the agent’s (2010, 143). It seems then that, without distinguishing the constitutive normativity of reasons from other normative dimensions in evaluating reasons, Kim’s account is committed to saying that we cannot consistently say: “Now I see why she did it, but it was something terrible to do.”

Consider again the starving girl. Most of us find her action inappropriate. Presumably, then, her deliberative principles or rules of decision are quite different from ours. If so, on Kim’s account, she and we are not in the same circle of rational intelligibility in Kim’s sense, which means that we can never understand her action. However, we perfectly well
understand why she acts as she does. Here is what my account says. Her action is intelligible, since we can simulate her, and she and we are in the same circle of Rationality. In fact, in order for us to assess an action as appropriate or inappropriate, the agent must be in our circle of Rationality; the action must be intelligible to us in the first place.

As usually understood, rules of deliberation and decision presuppose reasons; the point of deliberation is to compare and weigh various reasons we have. There may be a good sense in which those who share the same rules of deliberation and decision constitute a circle of “rationality.” In which case, however, the rationality involved will not be the kind of rationality that sets limits on the intelligibility of actions. Agents in the same circle of rationality in that sense will be those who generally agree about what is the rational thing to do on given occasions in some thick sense of ‘rational.’

On my account, by contrast, the rationality involved is the constitutive normativity of reasons. So what agents in the same circle of Rationality agree about is what could be a reason for a given action—if you and I don’t agree about this, our actions remain unintelligible to each other. And, in turn, what could be a reason is determined by what we can simulate—if you and I are different in this, we disagree about what could be a reason for an action. This means that on my approach, “principles of action” are no more than patterns of simulation. Agents in the same circle of Rationality share the same patterns of simulation.

In discussing the circle of rational intelligibility, Kim notes a resemblance between his result and Davidson’s thesis about radical interpretation, that we can radically interpret only subjects with whom we share most of our beliefs. He suggests that the similarity between them is not a mere coincidence but may well arise from “a shared conception of what it is for us to understand others—understand what they mean, what they believe and desire, and why they act the way they do” (2010, 146-7). I believe that a proper understanding of the insight Kim provides here will yield deeper understanding of fundamental issues in philosophy of mind and action. Now, it seems to me that the similarity Kim points to is more plausibly secured when we understand the notion of rationality involved in his circle
of rational intelligibility in the way I propose to understand it, that is, as the constitutive normativity of reasons. I also think that the resemblance becomes more conspicuous when we consider what Lewis (1974) calls the “rationalization principle” rather than Davidson’s charity principle. This principle says that the subject should be interpreted as a rational agent so that as many of his beliefs and desires as possible come out rational. If we add here my simulationist account of Rationality, it might be drawn that we can radically interpret only subjects with whom we share most of patterns of simulation.

8. Conclusion

The crucial task for anyone who advocates a normative model of action explanation is to show how a normative relation can underwrite an explanatory relation. I believe this can be done only if we discern the constitutive normativity of reasons; otherwise we cannot put to rest Hempel’s criticism that normativity of a reason is irrelevant to its explanatory force. My starting point is the observation that if an attitude is to serve as a reason for an action, it must be appropriately related to the action, in virtue of which the attitude counts as a reason for the action. And I have proposed to analyze this constitutive normative relation, which I call the Rationalizing relation, in terms of simulation.

The present approach is a synthesis of the ideas from Dray, Hempel, Davidson, and Kim. The resulting account is a Davidsonian, simulationist account of action explanation and rationality. In the literature, the debate on action explanation is usually characterized as a debate between two opposing camps. On the one side of the debate is the causalist camp to which Hempel and Davidson belong, and on the other side is the anti-causalist camp which includes Dray and neo-Wittgensteinians. If I am right, however, Dray’s and Davidson’s views can be coherently combined and considered a promising middle ground between Hempel and neo-Wittgensteinians.

The simulation theory has been primarily presented as an empirical hypothesis about how we interpret others, or what cognitive mechanism underlies our actual interpretive practice. Much of the philosophical interest that mental simulation has elicited, however,

13E.g., Ryle (1949), Anscombe (1957), Melden (1961), and von Wright (1971).
lies in the hope that understanding our interpretive practice can shed light on the very nature of mental states, namely on the constitutive issue of what it is for a person to have an attitude with a certain content. Our discussion points to one way this hope might be realized. If I am generally right that simulation can help capture the notion of rationality that governs propositional attitudes psychology, simulation may play a constitutive role in determining what attitudes a subject has, not just an epistemic or methodological role.
CHAPTER 3

Overdetermination, Counterfactuals, and Mental Causation

This chapter addresses the problem of mental causation, focusing on the so-called exclusion, or overdetermination, problem. In particular, I take up the question whether counterfactuals can yield an appropriate notion of causal redundancy, and examine how this issue bears on the mental causation debate. It is often suggested that the overdetermination argument relies on a certain ‘thick’ conception of causation, and that if causation were understood in terms of counterfactuals, the overdetermination problem simply would not arise. I try to motivate and give substance to this idea and then show why this strategy, though initially appealing, leaves the problem unresolved; rather, it only reveals the expressive poverty of counterfactual causal talk.

1. Introduction

Our actions are a product of our mental states; we do things because of what we want, believe, hope, fear, and so forth. Presumably, human agency requires the reality of mental causation. However, since Descartes, explaining how minds can manage to affect bodies has been a central problem in the metaphysics of mind. Indeed, there are many sources that generate difficulties for mental causation, and so there are many distinct problems of mental causation. In this chapter, we will focus on one of these problems, often called the exclusion, or overdetermination, problem.1

The overdetermination problem is most often presented as an argument against nonreductive physicalism. On this overdetermination argument, a nonreductive physicalist view of mind has the consequence that it makes every case of mental causation a case of overdetermination, but it is not plausible to postulate such causal redundancy as a common feature

1I will distinguish between the exclusion problem and the overdetermination problem in the next chapter. Here I use the terms interchangeably.
of the world. There are three main lines of nonreductivist response. The first is to concede that mental causation involves overdetermination but to argue that this is perfectly acceptable. The second is to deny that nonreductive physicalism is committed to the view that the effects of mental causes are always overdetermined. The third is to argue that the overdetermination problem turns on a particular, problematic conception of causation.

In pursuing these lines, nonreductivists often appeal to counterfactuals in one way or another. For example, in defending an overdetermination picture of mental causation, Jonathan Schaffer (2003, 2004) presents a counterfactual characterization of overdetermination and argues that our actions are indeed overdetermined by mental and physical causes. Karen Bennett (2003, 2008) relies on essentially the same counterfactual criteria for overdetermination to argue that nonreductive physicalism is not committed to mental-physical overdetermination. Barry Loewer (2003, 2007) also invokes counterfactuals but focuses rather on our conception of causation. He argues that the force of the overdetermination argument turns on a productive or generative conception of causation, but if causation is understood in terms of counterfactuals, concerns about overdetermination simply do not arise.

I find all these attempts less than satisfying, and in this chapter I show why. More broadly, the aim of this chapter is to assess the prospects for strategies that invoke counterfactuals to resolve the overdetermination problem. I take up the question whether counterfactuals can yield an appropriate notion of causal redundancy, and explore how this issue bears on the mental causation debate. The chapter divides into three main sections, which roughly correspond to the three lines of response to the overdetermination problem.

In section 2, I consider the dependent overdetermination approach to mental causation and identify the primary source of our metaphysical discomfort about overdetermination. I argue that as it stands, this approach leaves our worry about causal redundancy unresolved.

Section 3 addresses the question whether counterfactuals can capture something essential and distinctive about overdetermination, and argues for a negative answer. I consider David Lewis’s and Schaffer’s definitions of overdetermination and possible revisions of
them, and show that none of these is necessary or sufficient for overdetermination. I then observe that counterfactual analyses of overdetermination face the same kinds of problems as counterfactual analyses of causation, and show that the source of these problems is that counterfactual analyses of overdetermination—just as those of causation—cannot accommodate the intuitive idea that causation is an intrinsic relation between events.

Section 4 explores the implications of these observations for the problem of mental causation. I begin by trying to motivate and give substance to the claim that the threat of overdetermination does not arise on a dependency conception of causation. I then show why this idea, though initially appealing, does not address the real issue. As we will see, the idea derives its spurious plausibility from the fact that a dependency conception of cause is unable to make sense of our pre-theoretic idea of causal redundancy. This means that the overdetermination problem cannot even be coherently stated on a dependency conception of cause, revealing a serious expressive poverty of counterfactual causal talk. I close by briefly discussing a possible picture of mental causation that suggests itself in light of my discussion.

2. The Overdetermination Problem

Here is how the overdetermination problem is typically set up. Suppose that certain entities or properties do not reduce to their underlying, basal conditions—very often physical—from which they arise in some sense. Suppose further that the higher-level phenomena are causally efficacious. This would then seem to create a widespread and systematic causal overdetermination, since any effects caused by the higher-level phenomena would also be caused by their underlying bases.

This problem arises for antireductionists in a number of domains, but it has been most intensively discussed in the philosophy of mind. Jaegwon Kim, for example, has used a version of it, known as the “exclusion problem,” to forcefully pressure his fellow physicalists toward reductionism (see, among others, Kim 1998, 2005). The upshot of the exclusion

---

2It arises, for instance, for second-order functional properties (Block 1990), dispositional properties (Prior, Parrotter and Jackson 1982), and ordinary objects (Merricks 2001).
problem is that we are left with a dilemma between epiphenomenalism and overdetermination, unless we are prepared to accept some form of reductionism. Most philosophers don’t find the first horn, epiphenomenalism, to be an option; Jerry Fodor, for instance, says it would be “the end of the world” (1990, 156). They tend to find the other horn more manageable. Indeed, an approach that has become quite popular in recent years is to admit that every case of mental causation involves overdetermination, but then to argue that this is not problematic (see, e.g., Segal and Sober 1991, Mellor 1995, Mills 1996, Loewer 2002, Schaffer 2003, Sider 2003).

The idea is this. The kind of overdetermination we have in the case of mental causation is importantly different from the one we have in standard examples of overdetermination, like firing squad cases. In the case of mental causation, the two overdetermining causes, unlike two bullets that hit the victim simultaneously, are tightly related to each other. The mental cause is supposed to be dependent on the physical cause as a matter of necessity: metaphysical or at least nomological. Let us distinguish dependent overdetermination, in which there is a dependency relation, nomological or metaphysical, between the overdetermining causes, from independent overdetermination, in which there is no such connection. Then, it may seem plausible to say, with Barry Loewer (2002, 657), that although cases of independent overdetermination are indeed rare, this provides no reason to think that cases of dependent overdetermination are unusual. As Ted Sider (2003, 722) points out, while widespread independent overdetermination may require widespread coincidences, widespread dependent overdetermination need not, given the necessary truth governing the relation between the overdetermining causes. All that follows, Jonathan Schaffer (2003, 28) holds, would be that “one sort of overdetermination is indeed everywhere,” but this is unsurprising and perfectly acceptable, for it is due “not to improbable conspiracy or coincidence, but rather to everyday lawful correlation.”

However, we need to ask: Why isn’t massive dependent overdetermination problematic? Why should the tight relation between the overdetermining causes make the overdetermination less troublesome? One is likely to believe it less objectionable if one thinks
that the source of our worries about overdetermination is “because it is wrong, other things equal, to postulate coincidences” (Block 1990, 159; see also Burge 1993, 115). But, as I see it, the principal source of the worry is causal redundancy rather than coincidence. What I take to be at the heart of the problem is this. To insist that mental causes are genuinely efficacious and also distinct from physical causes, which makes the effect look overdetermined, appears to be to postulate global causal superfluousness, not global coincidences, as a common, objective feature of the world. This is metaphysically troubling. And this kind of worry is hardly allayed by showing that in cases of mental causation, unlike standard cases of overdetermination, there are lawful or metaphysical connections between the overdetermining causes. This of course shows that nonreductive physicalism is not committed to worrisome coincidences on a global scale. However, the idea that nature should contain such built-in, pervasive causal redundancies still strikes us as weird. Still, as Stephen Schiffer (1987, 148) puts it, “it is hard to believe that God is such a bad engineer.” In fact, I think systematic overdetermination might well be viewed as more objectionable than merely widespread overdetermination. Naturally designed or pre-established causal redundancy appears to be even harder to accept.

One might wonder whether my objection rests on a terminological choice. A simple way of putting my point is that dependent overdetermination looks problematic insofar as it is a kind of overdetermination. To this, it might be replied that dependent overdetermination is not a genuine form of overdetermination. Cases of what I have called ‘dependent overdetermination’ differ importantly from standard cases of overdetermination. Shouldn’t we say then that they do not in fact involve overdetermination at all? This prompts us to ask: What exactly is meant by overdetermination? Why doesn’t the sort of overdetermination (or just causation) involved in mental causation count as genuine overdetermination?

Karen Bennett (2003, 2008) tackles this issue when she argues that the nonreductive physicalist view of mental causation does not involve overdetermination. She admits that the burden is on the nonreductive physicalist here, noting that “she needs to be able to argue

---

3It may be worth here emphasizing that, as Kim (1998, 52) does, “It is important to see that the problem that we face arises because the two putative causes are not independent events.”
that the effects of mental causes are not overdetermined, and to explain why they are not” (2003, 474). So Bennett emphasizes the need for “a well-motivated and non-handwaving way to break the analogy” between mental causation cases and other paradigmatic cases of overdetermination (2003, 492).

It may, then, be better put my point this way. Neither simply breaking the analogy between the two types of overdetermination nor merely showing that the kind of overdetermination (or causation) involved in mental causation doesn’t require coincidence will suffice for solving the problem. What needs to be shown is that the kind of overdetermination (or causation) involved in mental causation is not really committed to what we find objectionable in massive overdetermination of the other kind. Now, what we find most objectionable in massive independent overdetermination is pervasive causal redundancy. Then, it doesn’t seem to me that either simple appeal to the tight connection between overdetermining causes or any other way of just breaking the analogy between the two types of overdetermination do much to help in this regard.

An example might be useful. To borrow a well-known case from Fred Dretske, a soprano sings, say, the word “shatter” at a high pitch, causing a nearby glass to break (1988, 79). In this case, the sound’s meaning what it means is causally irrelevant to its effect on the glass; what is causally responsible for the shattering is the sound’s acoustic or physical properties. Now suppose, however, that the soprano turns out to be a sorceress and her casting a spell by uttering “shatter” also caused the glass to shatter—if she had uttered instead the word “freeze,” by contrast, the glass would have frozen. The sorceress, thus, didn’t need to actually sing; whispering “shatter” would have sufficed to shatter the glass. In this example, the shattering of the glass seems to be causally overdetermined, as not only the total acoustic properties but also the semantic property of the sound are each causally sufficient for the shattering. Now, presumably, a sound’s semantic property is in some way related to, or dependent upon, its acoustic properties, in a way that the bullets’ momentums in the firing squad are not. This example therefore looks importantly different from standard cases of overdetermination, in that there is a tight connection between the
overdetermining causes. However, it nonetheless seems as though the example certainly involves causal redundancy. And no one would think that such an example serves as a model for mental causation, though it looks like a case of dependent overdetermination par excellence. Indeed, the overdetermination involved in this example, and the idea of dependent overdetermination itself, sound to me, in Jackson’s phrase, like a “curious and ontologically extravagant kind of overdetermination” (1998, 92).

In fairness to Bennett, she does not simply try to break the analogy. Rather, she tries to show that mental causation, or dependent overdetermination, is not really overdetermination, looking for principled criteria for overdetermination. But is there a serviceably clear definition or characterization of what overdetermination is? No doubt, the answers to the questions of whether nonreductive physicalism is really committed to massive overdetermination, and how objectionable the commitment is, depend on exactly how overdetermination is understood.

It seems quite widely believed that counterfactuals are crucial to the very understanding, or capturing, of the intuitive notion of overdetermination. This is no surprise given the popularity of counterfactual approaches to causation, though analyzing causation in terms of counterfactuals does not commit one to analyzing overdetermination in terms of counterfactuals as well, nor vice versa. The guiding idea, which can be found in David Lewis (1986a, 2000), is this: an event is overdetermined when it has two distinct causes such that if either one of them had occurred without the other, the effect would still have occurred. Something like this is often assumed, explicitly or implicitly, in the debate over the overdetermination problem. Schaffer, for example, presents a definition of overdetermination along this line and claims that “If the two-levels case meets the requirements for overdetermination, then it is overdetermination, end of story” (2003, 28). Bennett also explicitly relies on the same basic idea when she argues that mental causation does not involve overdetermination because it does not meet some counterfactual-based criteria she takes to be necessary for overdetermination. Evaluating the prospects for this idea is the task of the next section.
3. Overdetermination and Counterfactuals

‘Causal overdetermination’ is a term of art and philosophers most usually rely on examples to illustrate the notion. Here is a textbook example. Suzy and Billy both throw rocks at a bottle. Each rock is perfectly aimed and thrown with sufficient force. The rocks strike the bottle at the same time and it shatters. It is said then that the shattering is causally overdetermined by Suzy’s and Billy’s rocks, for one alone would have shattered the bottle. A question can be raised here whether counterfactuals can help us go beyond examples and give a workable general definition or analysis of overdetermination. Before we move on, though, let us briefly address a concern about the very notion of overdetermination.

Here is the concern. Everyone will agree that there is something unusual about cases like Suzy and Billy. Presumably, the peculiarity involved is what makes us judge that such cases of overdetermination would be quite rare, if not impossible, and cannot serve as a proper model for mental causation. However, it is not entirely clear precisely what the peculiarity consists in. It seems not altogether implausible, for example, to say that the peculiarity consists merely in coincidence, so that from a metaphysical or causal point of view there is nothing especially peculiar or problematic about such cases of overdetermination. What is problematic is rather our intuitive notion of overdetermination. It might turn out that there is an incoherence in the concept of overdetermination. On this line of thinking, there is no genuine causal redundancy in nature; it might be just in the eye of beholder.

I find this sort of line interesting and challenging. However, even if its underlying thought is well taken, it does not readily make the overdetermination problem go away. Let me first be clear that the thought is not just that there are actually no true instances of overdetermination. Martin Bunzl (1979) has argued that every alleged case of overdetermination will resolve into either a case of causal preemption or of joint causation. Presumably, each rock in the case of Suzy and Billy would have made some minute difference to the time and manner of the shattering of the bottle (when an appropriately modally fragile event is assigned to the shattering). Admittedly, given the actual laws of nature, we cannot simply
stipulate that the shattering would occur precisely the same way whether one rock shatters it or two. It might be then that overdetermination is not nomologically possible, so that any true example of overdetermination must necessarily involve a fairy-tale device as in the soprano-sorceress case. However, it should be noted that, if overdetermination is nomologically impossible, this only shows that on nonreductive physicalism, mental causation too might involve phenomena of the sort that are not nomologically possible. The force of the overdetermination problem, thus, does not turn on there being realistic examples of overdetermination.

The thought, however, may be that the notion of overdetermination is not merely empirically implausible, but rather conceptually problematic. Indeed, there seems something puzzling about the notion of overdetermination. How could it be that each overdetermining cause brings about the effect all on its own? In response, I would point out that there is anyway a clear difference, at least on an intuitive level, between the putative examples of overdetermination and those that are not. Suppose, for instance, two wizards cast spells on a prince at the same time and the prince is turned into a frog. It will make perfect sense to ask whether each of the spells is sufficient on its own or just necessary to turn the prince into a frog. It seems clear that there is a significant difference between those two possibilities, a difference concerning causal redundancy. If the two possibilities are genuine metaphysical possibilities and it makes sense to ask about the difference between them, then I think this is enough for the worry over causal redundancy to arise. For it would still appear problematic if the case of mental causation, under nonreductive physicalism, looks like the version of the two wizard case where each spell suffices to turn the prince into a frog.

### 3.1. Counterfactual Analyses of Overdetermination

One common, natural way of characterizing overdetermination is to say that an event is overdetermined when it has two distinct causes such that each one without the other would still have caused the event. Let us begin with a simple formulation of this idea, which we may call the ‘simple counterfactual analysis’ (SCA):
SCA: \( c_1 \) and \( c_2 \) causally overdetermine \( e \) if and only if

(O1) both \( c_1 \) and \( c_2 \) are causes of \( e \);

(O2) if \( c_1 \) had occurred without \( c_2 \), \( e \) would still have occurred \((c_1 \& \neg c_2 \rightarrow e)\);

(O3) if \( c_2 \) had occurred without \( c_1 \), \( e \) would still have occurred \((\neg c_1 \& c_2 \rightarrow e)\).

As it stands, the analysis presupposes no particular account of causation. The ‘cause’ in (O1) need not necessarily be understood in terms of counterfactuals; if you have any favored theory of causation, plug it into SCA and you will get your own version of this counterfactual analysis of overdetermination.\(^4\)

To be sure, any adequate analysis of overdetermination should be able to distinguish overdetermination from preemption and joint causation. In SCA, (O1) purports to distinguish overdetermination from preemption, while (O2) and (O3) purport to distinguish overdetermination from joint causation. Brief reflection, however, shows that SCA cannot handle preemption cases adequately. The problem is that events along a single causal chain easily pass SCA when there is a preempted backup process. Consider a canonical example of preemption: Billy and Charlie both throw rocks at a window; but Billy’s rock arrives first and shatters the window. The causal structure of this case can be represented by a simple, Lewisian “neuron diagram”:

![Figure 1](image)

Here \( c_1 \), the firing of neuron \( c_1 \), represents Billy’s throw; \( d_1 \) Charlie’s throw; \( c_2 \) the collision between Billy’s rock and the window; \( d_2 \) the collision between Charlie’s rock and the window; \( e \) the shattering of the window.\(^5\) SCA delivers the right verdict that \( c_1 \) and \( d_1 \)

\(^4\) I focus on cases with just two overdetermining causes. But the analysis can be extended to cover cases with three or more overdetermining causes, for example, in the way that Schaffer (2003) does.

\(^5\) Diagram convention: bold letters name neurons, italicized letters the events of their firing; filled circles represent neurons that fire, unfilled circles neurons that do not fire; arrows represent stimulatory connections, blobs inhibitory ones; solid lines represent actualized connections, and dashed lines unactualized ones.
do not overdetermine \( e \), for \( d1 \) is not a cause of \( e \). The problem, however, arises when we look at the events \( c1, c2, \) and \( e \). Both \( c1 \) and \( c2 \) are clearly causes of \( e \); moreover, if \( c2 \) had occurred without \( c1 \), \( e \) would still have occurred; if \( c1 \) had occurred without \( c2 \), \( e \) would still have occurred thanks to the process initiated by \( d1 \). Thus, the conditions (O1)-(O3) are all satisfied by \( \langle c1, c2, e \rangle \) and, therefore, we are forced to say that \( c1 \) and \( c2 \) overdetermine \( e \). But \( e \) is not overdetermined, and a fortiori not by \( c1 \) and \( c2 \).

What’s wrong with SCA? It seems we usually construe overdetermination so that if either of the overdetermining causes had not occurred, the effect would have been caused \textit{by the other overdetermining cause}. Recall the case of Suzy and Billy. If Suzy’s throw, one of the overdetermining causes, had not occurred, the shattering would have been caused by Billy’s throw, the other overdetermining cause. However, in the case of Figure 1, if \( c1 \) had occurred without \( c2 \), \( e \) would have been caused not by \( c1 \) but by the backup process initiated by \( d1 \). In fact, even if neither \( c1 \) nor \( c2 \) had occurred, \( e \) would have occurred due to the backup. So the problem is that the conditions (O2) and (O3) together are too weak; they are not sufficient to guarantee that in the absence of one overdetermining cause, the other overdetermining cause is that which now causes the effect.

This diagnosis suggests repair strategies. In fact, Lewis’s and Schaffer’s respective analyses can be regarded as two ways of revising SCA on the basis of this diagnosis. First, Schaffer’s analysis of overdetermination (2003, 23), SAO:

\textbf{SAO:} \( c1 \) and \( c2 \) causally overdetermine \( e \) if and only if

(O1) both \( c1 \) and \( c2 \) are causes of \( e \);
(O2*) if \( c1 \) had occurred without \( c2 \), \( c1 \) would still have \textit{caused} \( e \);
(O3*) if \( c2 \) had occurred without \( c1 \), \( c2 \) would still have \textit{caused} \( e \).

SAO tightens up SCA by requiring that in the absence of one of the overdetermining causes, the effect should not simply have occurred, but should have been \textit{caused} by the other overdetermining cause.
Next, Lewis’s analysis of overdetermination, LAO, goes this way.⁶

**LAO:** *c*₁ and *c*₂ causally overdetermine *e* if and only if

1. (O1) both *c*₁ and *c*₂ are causes of *e*;
2. (O2) if *c*₁ had occurred without *c*₂, *e* would still have occurred (*c*₁&¬*c*₂ □→ *e*);
3. (O3) if *c*₂ had occurred without *c*₁, *e* would still have occurred (¬*c*₁&*c*₂ □→ *e*);
4. (O4) if neither *c*₁ nor *c*₂ had occurred, *e* would not have occurred (¬*c*₁&¬*c*₂ □→ ¬*e*).

LAO explicitly strengthens SCA by adding the fourth condition (O4) while preserving (O1)-(O3) intact. Here adding (O4) is expected to yield a similar result as tightening up (O2) and (O3) into (O2*) and (O3*) did for SAO, but without building in extra causal terms.

It is easily seen now that each analysis gets the foregoing example right. SAO does not say that *c*₁ and *c*₂ in Figure 1 overdetermine *e*, because (O2*) is not satisfied. If *c*₁ had occurred without *c*₂, *c*₁ would not have caused *e*; instead, *e* would have been caused by the process initiated by *d*₁. LAO also does not count *c*₁ and *c*₂ as overdetermining causes of *e*, because (O4) is not satisfied. Even if neither *c*₁ nor *c*₂ had occurred, *e* would still have occurred thanks to *d*₁.

Both analyses may then appear to nicely address the shortcoming of SCA. However, a slight variation on the case of Billy and Charlie shows that they do not. Consider the following:

Here neuron *c*₀ fires and sends stimulatory signals to *c*₁ and *d*₁. For instance, *c*₀ represents someone’s ordering Billy and Charlie to throw their rocks at the window. The problem arises when we look at the events *c*₀, *c*₁, and *e*. ⟨*c*₀, *c*₁, *e*⟩ meets all of the conditions of

⁶To be precise, Lewis does not officially put forward an analysis or definition of overdetermination. Something like LAO, however, is suggested when he gives a definition of redundant causation. See Lewis 1986a, 193; 2000, 80.
SAO and LAO. First, the event $e$ has two causes, $c0$ and $c1$, such that if $c1$ had occurred without $c0$, $c1$ would still have caused $e$; and if $c0$ had occurred without $c1$, $c0$ would still have caused $e$. Thus, it comes out on SAO that $e$ is overdetermined by $c0$ and $c1$. We also see that if neither of the causes $c0$ and $c1$ had occurred, $e$ would have not occurred either. Thus, LAO gives the same verdict that $c0$ and $c1$ overdetermine $e$. But it is clear that $e$ is not overdetermined, and a fortiori not by $c0$ and $c1$. Thus, neither SAO nor LAO provides a sufficient condition for overdetermination.

3.2. Counterfactual Analyses Strengthened: Still Insufficient. One might point out that in the foregoing example $c0$ is a cause of $c1$, so that they are not causally independent of each other, whereas in standard examples of overdetermination, the overdetermining causes are supposed to be causally independent of each other (as they are in the case of Suzy and Billy). This suggests that SAO and LAO might be simply amended by requiring further that one overdetermining cause be causally independent of the other, however this may be elaborated. We would then have the following:

**SAO*: $c1$ and $c2$ causally overdetermine $e$ if and only if

(O1*) $c1$ and $c2$ are causes of $e$ and causally independent of each other;
(O2*) if $c1$ had occurred without $c2$, $c1$ would still have caused $e$;
(O3*) if $c2$ had occurred without $c1$, $c2$ would still have caused $e$.

**LAO*: $c1$ and $c2$ causally overdetermine $e$ if and only if

(O1*) $c1$ and $c2$ are causes of $e$ and causally independent of each other;
(O2) if $c1$ had occurred without $c2$, $e$ would still have occurred ($c1 \& \neg c2 \rightarrow e$);
(O3) if $c2$ had occurred without $c1$, $e$ would still have occurred ($\neg c1 \& c2 \rightarrow e$);
(O4) if neither $c1$ nor $c2$ had occurred, $e$ would not have occurred ($\neg c1 \& \neg c2 \rightarrow \neg e$).

Both SAO* and LAO* might seem strong enough to capture the intuitive idea of overdetermination. However, as I will now show, this kind of move does not work either.
Let us first examine why SAO* is not sufficient for overdetermination. Consider first a case of joint causation: This time Suzy and Billy fire guns at the legs of a chair; their bullets hit two legs, say, the first and second leg respectively; the two legs get broken; the chair loses its balance and falls over. Suppose here that Suzy’s gunshot, \(c1\), and Billy’s gunshot, \(d1\), jointly cause the chair’s falling, \(e\)—one leg’s breaking alone would not make the chair fall over. Now, suppose further that Sally and Wiley were also actually in the vicinity and they too fired guns at the same legs of the chair, almost at the same time that Suzy and Billy shot. But Suzy’s and Billy’s bullets hit the legs first. Sally’s and Wiley’s bullets pass through the spaces where the first and second leg had been a moment before, and the chair falls over. The following diagram represents this case:

Here \(c1\) represents Suzy’s gunshot; \(d1\) Billy’s; \(c1^*\) Sally’s; and \(d1^*\) Wiley’s; \(c2\) represents the hitting of the first leg by Suzy’s bullet; \(d2\) the hitting of the second leg by Billy’s; \(c2^*\) the hitting of the first leg by Sally’s; \(d2^*\) the hitting of the second leg by Wiley’s; and \(e\) represents the chair’s falling. This is a case of preemption: \(c1\) is a preempting cause and \(c1^*\) is a preempted backup for the breaking of the first leg; \(d1\) is a preempting cause and \(d1^*\) is a preempted backup for the breaking of the second leg. It is clear that this is not a case of overdetermination. Suzy’s and Billy’s gunshots do not overdetermine the chair’s falling; they just jointly cause it.

However, \(\langle c1, d1, e \rangle\) satisfies all of the conditions of SAO*. Clearly, \(c1\) and \(d1\) are causes of \(e\) and causally independent of each other. And both counterfactuals (O2*) and (O3*) are true. If \(c1\) had occurred without \(d1\), then \(c1\) would still have caused \(e\); in the situation where Suzy alone fires her gun without Billy, Suzy’s gunshot would still have been a cause of the chair’s falling, for, in combination with Wiley’s gunshot, it would
still have caused the chair to fall over. And similarly, if \( d1 \) had occurred without \( c1 \), \( d1 \) would still have caused \( e \). Thus, on SAO*, \( c1 \) and \( d1 \) overdetermine \( e \). But surely \( e \) is not overdetermined, and \textit{a fortiori} not by \( c1 \) and \( d1 \).

It might be objected that it is questionable whether (O2*) and (O3*) are really true in this example. Consider the situation where \( c1 \) occurs without \( d1 \). Suzy alone fires her gun without Billy and her bullet hits the chair’s first leg; a split second later, Sally’s bullet passes through where that first leg had been and Wiley’s bullet hits the chair’s second leg, thereby causing the chair to fall over. Then, is it still right to say that in this situation Suzy’s gunshot caused the chair’s falling? It seems not in some sense. It appears natural to say rather that if Suzy alone had fired without Billy, then Suzy’s bullet would have merely caused the chair’s \textit{losing one leg} rather than its \textit{falling}; it is Wiley’s bullet that would have caused the chair’s \textit{falling} a split second later.

This objection is based on the observation that in the situation where Suzy’s gunshot occurs without Billy’s, there is a temporal gap between the time Suzy’s bullet hits the chair’s first leg and the time Wiley’s bullet hits the chair’s second leg. But this is not an essential feature of the example. Suppose that Suzy’s, Billy’s, Sally’s, and Wiley’s gunshots all hit the chair’s legs at the same time. In this case, there seems genuine causal \textit{redundancy}; the chair’s falling, indeed, seems overdetermined. But we need to be cautious about which events overdetermine which effects. What seems clear about this case is this: Suzy’s and Sally’s shots, \( c1 \) and \( c1^* \), overdetermine the \textit{breaking of the first leg} of the chair; Billy’s and Wiley’s shots, \( d1 \) and \( d1^* \), overdetermine the \textit{breaking of the second leg} of the chair. Now, is it right to say that Suzy’s and Billy’s shots, \( c1 \) and \( d1 \), overdetermine the \textit{chair’s falling}? Intuitively, it seems not; each of the gunshots makes its own causal contribution to the chair’s falling by \textit{breaking one leg} of the chair, and so needs the other’s help in order to bring about its falling. The right thing to say seems to be rather this. It is the events \( c1 \)-\textit{cum-}\( d1 \) and \( c1^*\)-\textit{cum-}\( d1^* \), not just \( c1 \) and \( d1 \), that overdetermine the \textit{chair’s falling}. That is, \( c1 \) and \( d1 \) \textit{jointly} cause it; Suzy’s and Billy’s gunshots “work together” to bring about
the chair’s falling. We can say, then, that \( e \) is indeed overdetermined, but not by \( c1 \) and \( d1 \). Thus, it remains true that SAO* is not sufficient for overdetermination.

Let us now turn to LAO*. Unlike SAO*, LAO* delivers the right verdict that \( c1 \) and \( d1 \) do not overdetermine \( e \) in the foregoing example. This is because \( \langle c1, d1, e \rangle \) does not satisfy the condition (O4); even if neither \( c1 \) nor \( d1 \) had occurred, \( e \) would still have occurred, thanks to \( c1^* \) and \( d1^* \). However, a slight variant on the story shows that LAO* and SAO* are nevertheless in the same boat. Suppose that Charlie was actually sitting on the target chair. Charlie observes Sally and Wiley aiming guns at the legs of his chair and he makes up his mind to block their shots, which is easy for him to accomplish somehow. At that time, however, he also notices two other gunmen, Suzy and Billy, aiming guns in his direction. Startled, Charlie puts his hands up and so does nothing to abort Sally’s or Wiley’s shots. As a result, the four shooters all fire and the chair falls over. The following represents this case:

![Figure 4](image)

Here, each process initiated by \( c1, d1, c1^* \), and \( d1^* \) represents the process initiated by Suzy’s, Billy’s, Sally’s, and Wiley’s aiming guns respectively. In addition, there is a process initiated by Charlie, the process from \( f1 \): \( f1 \) represents Charlie’s noticing Suzy and Billy aiming guns; \( f2 \) Charlie’s decision to block Sally and Wiley’s shots; \( f3 \) Charlie’s blocking their shots. But the stimulatory signal that \( f1 \) sends to \( f2 \) is blocked by inhibitory signals from \( c1 \) and \( d1 \). However, if \( f2 \) had fired, it would have caused \( f3 \) to fire, which in turn would have aborted the processes from \( c1^* \) and \( d1^* \). In this case, \( \langle c1, d1, e \rangle \) meets all of the conditions of LAO*—not only (O1*)-(O3) but also (O4); if neither \( c1 \) nor \( d1 \) had occurred, then \( e \) would not have occurred. But Suzy’s and Billy’s shots do not overdetermine the chair’s falling. Thus, LAO* is not sufficient for overdetermination either.
3.3. Even SCA Not Necessary for Overdetermination. One might think that although the counterfactual analyses we have considered do not provide sufficient conditions for overdetermination, their claim to necessity may be on a firmer footing. The thought is that counterfactuals can catch something at the very core of the notion of overdetermination, so that the truth of certain counterfactuals such as (O2) and (O3) is at least necessary for overdetermination. That is, when an event is overdetermined by two causes, something like the following must be true: if either one of the causes had occurred without the other, the effect would still have occurred. As we will see, however, the truth of (O2) or (O3) is not necessary for overdetermination. This means that even SCA, the weakest form of the counterfactual analyses, fails to provide a necessary condition for overdetermination.

Consider again the original case of Suzy and Billy, in which their rocks strike a bottle at the same time. Now suppose that Sally, the owner of the bottle, was also actually in the vicinity. Sally notices Suzy throwing a rock at the bottle and is about to block Suzy’s rock. But she also notices another rock being thrown by Billy, her beloved son. Sally doesn’t want Billy to be disappointed, so she does nothing. As chance has it, both throws are perfectly accurate; the bottle shatters and Billy hurrahs. However, if Suzy alone had thrown her rock without Billy, Sally would have intervened, and the bottle would not have shattered. The following represents this case:

Here each process initiated by $c1$ and $d1$ represents the process initiated respectively by Suzy’s and Billy’s throw; $e$ represents the bottle’s shattering. In addition, there is a process initiated by $f1$: $f1$ represents Sally’s noticing Suzy’s throw; $f2$ Sally’s decision to block Suzy’s rock; $f3$ Sally’s extending her hand to block Suzy’s rock. In this case, if $c1$ had occurred without $d1$, $e$ would not have occurred. Yet it still seems clear that the shattering
is overdetermined; the bottle gets struck by two rocks anyway. But the counterfactual ‘if $c_1$ had occurred without $d_1$, $e$ would still have occurred’ is false.

There may be an objection along the following lines. It might be granted that $e$ is in fact overdetermined. But, again, we need to be cautious about which events overdetermine the effect. Let us first focus our attention on $c_1$, Suzy’s throwing her rock. In our scenario, Suzy’s throw really looks redundant with respect to bringing about the shattering. Even if $c_1$ had not occurred, $d_1$, Billy’s throw, by itself would have caused $e$ without any interference. Let us now focus on $d_1$ and consider the situation where Billy’s throw had not occurred. Without $d_1$, $c_1$ alone would not have brought about $e$, since the process from Suzy’s throw would have been cut off by Sally’s interruption before causing the shattering. Then, Billy’s throw, unlike Suzy’s throw, seems to be indispensable to bringing about the shattering, and so it seems that $d_1$ is not redundant with respect to bringing about $e$. This observation may lead us to retract our initial judgment that $d_1$ is an overdetermining cause of $e$. At least, the kind of asymmetry we note between $c_1$ and $d_1$ suggests that the situation here is different from standard cases of overdetermination. In fact, the right thing to say seems to be this. It is not $c_1$ and $d_1$ but $c_3$ and $d_3$, the hittings of the bottle by Suzy’s and Billy’s rocks, that overdetermine the shattering. And the fact that $c_3$ and $d_3$ overdetermine $e$ misled us into forming the intuition that it is overdetermined by $c_1$ and $d_1$.

Such an asymmetry between the causes, however, is not an essential feature of the example, nor is the fact that it involves cutting off a process leading from one of the causes to the effect when the other cause had not occurred. To see this, consider the following example. Two sergeants give orders to a platoon of soldiers. Both the sergeants shout ‘Advance!’ at the same time, and the soldiers move forward. Now, suppose further that a major was standing there too. He was actually about to order the soldiers to retreat. But the major hears the two sergeants giving the orders to advance, so he does nothing. However, if only one of the sergeants had ordered the advance, he would have ordered the soldiers to retreat, which would have caused them to do so. The following represents this case:
Here, $c_1$ and $d_1$ represent the two sergeants’ orders to advance; $g_1$ represents the major’s watching; $g_2$ the major’s order to retreat; and $e$ the soldiers’ advancing. The inhibitory signals from $c_1$ and $d_1$ prevent $g_2$ from firing; without $c_1$ or $d_1$, $g_2$ would have occurred. (So they jointly prevent $g_2$; that is, if one of $c_1$ and $d_1$ had not occurred, the other alone would have failed to prevent $g_2$ from occurring.) And if $g_2$ had occurred, it would have prevented $e$ from occurring, without involving any cutting-off of a process from $c_1$ or $d_1$ on its way to $e$. In this case, it still seems true that the two sergeants’ orders overdetermine the soldiers’ advancing; the major just does nothing. But both counterfactuals (O2) and (O3) are false: if $c_1$ had occurred without $d_1$, $e$ would not have occurred; likewise, if $d_1$ had occurred without $c_1$, $e$ would not have occurred. Thus, SCA fails to provide a necessary condition for overdetermination. Then the strengthened versions of SCA, i.e., SAO, LAO, SAO*, and LAO* are a fortiori not necessary for overdetermination.

3.4. Diagnosis. We have seen that the counterfactual analyses thus far considered provide neither necessary nor sufficient conditions for overdetermination. What is the source of troubles for the idea of analyzing overdetermination in terms of counterfactuals? First observe that the counterexamples we have discussed involve the phenomena of preemption or double prevention. These are two much-discussed phenomena which have proved to be serious difficulties for counterfactual analyses of causation. In fact, it can be seen that the counterfactual approach to overdetermination faces problems that are of the same kinds as those facing the counterfactual approach to causation. Let us first briefly reflect on how cases of preemption and double prevention make trouble for the counterfactual approach to causation.
First, to see the issue with causal preemption, start by considering a situation where Billy alone throws a rock at a window and shatters it. Here Billy’s throwing his rock is a cause of the shattering and, clearly, the shattering counterfactually depends (or just ‘depends’ for short) on Billy’s throw; if Billy had not thrown his rock, the shattering would not have occurred. However, by introducing some extra backup process, we can make it the case that the shattering does not depend on Billy’s throw. The case of Billy and Charlie, which is represented in Figure 1, is one such example of preemption. In this case, it is clear that Billy’s throw, not Charlie’s, caused the shattering, but there is no counterfactual dependence between Billy’s throw and the shattering; if Billy’s throw had not occurred, the shattering still would have occurred, thanks to Charlie’s throw. It seems then that counterfactual dependence cannot be a necessary condition for causation.

Next, to see the difficulty double prevention poses, start by imagining any two events that seem to be utterly causally irrelevant to each other. Consider, for example, Billy’s throwing his rock toward the bottle and Suzy’s rock reaching the bottle, in the original case of Suzy and Billy. Here, of course, there is no counterfactual dependence between these two events. However, we can make them exhibit proper dependence by introducing some changes on external goings-on. The modified case of Suzy, Billy, and Sally, which is represented in Figure 5, is one such example involving double prevention. In this case, Sally sees Suzy throwing her rock at the bottle but does not block it because she also sees Billy, her son, throwing his rock. Here Billy’s throwing his rock prevents Sally from blocking Suzy’s rock which, had it not been prevented, would have prevented Suzy’s rock from reaching the bottle (hence the name of ‘double prevention’). In this situation, the event of Suzy’s rock reaching the bottle straightforwardly depends on Billy’s throwing his rock; if Billy had not thrown his rock, Suzy’s rock would not have reached the bottle. Yet, intuitively, it does not seem to be correct to say that Billy’s throwing his rock caused Suzy’s rock to get all the way to the bottle. Here the guiding intuition is that Sally’s just keeping her eyes on the circumstances should not alter our initial judgment that the flight of Suzy’s
rock to the bottle was caused not by Billy’s throwing his rock but by Suzy’s throwing the rock. If this is right, counterfactual dependence is not sufficient for causation.

It can now be seen how the counterfactual approaches to causation and overdetermination face the same kinds of difficulties, in fact ones that stem from the same source. The common root of the difficulties is that counterfactual dependency is an extrinsic matter, while there is a strong intuition that causation is an intrinsic relation. As the above cases of preemption and double prevention illustrate, whether an event depends on another is affected by extrinsic changes in the circumstances. Consider the case of Billy and Charlie, and compare the situation with that in which Billy alone throws his rock without Charlie. The shattering of the window depends on Billy’s throw in the latter situation, but not in the former. Similarly, consider the case of Suzy, Billy, and Sally, and compare the situation with that in which Sally is absent. Suzy’s rock reaching the bottle does depend on Billy’s throwing his rock in the former situation, but not in the latter. In this sense counterfactual dependence is not an intrinsic matter.

By contrast, there seems a strong intuitive appeal to the idea that causation is an intrinsic relation. The idea is that causation is a relation between events that is instantiated just in virtue of the intrinsic properties of, and relations between, the events.⁷ Indeed, it is this idea that guides the intuitive judgment that in the case of Billy and Charlie, the causal features of the process from Billy’s throw to the shattering should remain the same regardless of whether the process from Charlie’s throw, which is extrinsic to the process from Billy’s throw, is added or not. And similarly for the judgment that in the case of Suzy, Billy, and Sally, the causal status of Billy’s throw with respect to the flight of Suzy’s rock should remain the same regardless of whether the story about Sally’s unactualized intervention is added or not. So the upshot is this: The problems for the counterfactual approach to causation arise from the conflict between the extrinsic character of counterfactual dependence and the strong intuition that causation is an intrinsic matter, and the cases of preemption and double prevention just highlight this conflict.⁸

⁷See Hall (2004b) for a more elaborated formulation and useful discussion of this idea.
⁸This observation is largely due to Ned Hall. See Hall (2004a).
It is not difficult now to see how this conflict raises the same kind of problems for the counterfactual approach to overdetermination. Consider first how the sufficiency of the counterfactual analyses has been shown to fail. We start with a case of joint causation, like the example in which Suzy’s and Billy’s shots jointly cause the chair’s falling. This case does not meet the conditions for overdetermination required by the counterfactual analyses, so the analyses do give the correct verdict that Suzy’s and Billy’s shots don’t overdetermine the chair’s falling. Now modify the case by introducing some extrinsic changes, for instance, by adding preempted backups of Sally’s and Wiley’s shots, as in Figure 3. The modified case then satisfies all of the counterfactual conditions of the analyses so that the analyses now say that Suzy’s and Billy’s shots overdetermine the chair’s falling. Yet, it seems, introducing such extrinsic changes as preempted backups does not affect our initial judgment that Suzy’s and Billy’s shots jointly cause the chair’s falling; their gunshots “work together” to bring about the falling. So the counterfactual conditions are not sufficient for overdetermination.

Consider now how the necessity of the analyses has been shown to fail. We start with a case of overdetermination, like the original case of Suzy and Billy. This case meets all of the conditions for overdetermination required by the counterfactual analyses, so the analyses initially give the correct verdict that Suzy’s and Billy’s throws overdetermine the bottle’s shattering. Now modify the case by introducing some extrinsic changes, for instance, by adding the story about Sally’s unactualized intervention involving double prevention, as in Figure 5. The modified case then no longer meets the counterfactual conditions of the analyses so that the analyses say that Suzy’s and Billy’s throws do not overdetermine the shattering. Yet, it seems, introducing such extrinsic changes does not affect our initial judgment that Suzy’s and Billy’s throws do overdetermine the shattering; the bottle gets struck by their two rocks anyway. So the counterfactual conditions are not necessary for overdetermination.

The upshot, then, is this. The counterfactual approach to overdetermination faces the same kinds of problems as the counterfactual approach to causation, and the common
source of all these problems is that counterfactuals cannot accommodate the idea that causation is an intrinsic relation.

4. Counterfactuals and Mental Causation

We have seen the troubles, and their source, for the idea of analyzing overdetermination in terms of counterfactuals. What lessons, then, can we draw from this observation for the problem of mental causation?

One thing that our discussion reveals is that Bennett’s and similar attempts to resolve the exclusion problem by invoking counterfactuals simply rest on a false premise. In trying to defuse the threat of overdetermination, Bennett (2003, 2008) argues that the truth of (O2) and (O3) is necessary for overdetermination, then attempts to show that the relevant mental/physical counterfactuals do not turn out to be true under nonreductive physicalism, so that mental causation does not involve overdetermination. Bennett’s argument thus hinges crucially on the idea that certain counterfactuals provide a necessary condition for overdetermination. However, if our considerations are on target, overdetermination does not seem to require any such counterfactual to be true.

Let us consider another kind of strategy that appeals to counterfactuals in dealing with the exclusion problem. It is often suggested that the overdetermination threat does not arise in the first place if causation is understood in terms of counterfactuals. The thought is that the intuitive force of the exclusion problem turns on a certain thick conception of causation that involves some kind of ‘causal oomph’ or ‘biff,’ as it is sometimes put, but that the problem can be avoided if we base our causal claims on counterfactual dependencies (Loewer 2002, 2007; see also Burge 1993).

This sort of deflating strategy has some prima facie appeal, which I think comes from the spurious plausibility of the idea thought that the notion of overdetermination presupposes a productive or generative conception of cause. As we will see, however, once we try to spell out the idea, the strategy only manages to avoid the problem by ignoring it. It might well be that the worry about overdetermination does not arise if causation is stipulatively
understood as counterfactual dependency. However, this is because counterfactual dependency cannot even make sense of the intuitive, pre-theoretic idea of causal redundancy that lies behind the notion of overdetermination. But merely moving to a conception of cause that does not even permit the problem to be formulated is not enough in itself to resolve the problem. My aim in the rest of the chapter is to make a case for this point.

4.1. Two Concepts of Causation and the Overdetermination Problem. The counterfactual approach to causation is one of the most favored approaches to causation. As we have seen, however, there are serious difficulties for the counterfactual approach, which are highlighted by preemption and double prevention cases. One might be inclined to think that what the troubles show is simply that the idea of analyzing causation in terms of counterfactuals is just hopeless—this will seem especially attractive to those who already believe that causation ought to be conceived as an intrinsic relation between events. If this is the proper lesson, and we can really show that causation is not analyzable in terms of counterfactuals, it will easily undercut any response to the overdetermination problem that rests on understanding causation as counterfactual dependence.

However, it may be premature to discard the idea of causation as counterfactual dependence altogether. Certainly, there is more to be said for the idea that counterfactual dependence captures something true and important about causation. Consider again the case of Suzy, Billy, and Sally depicted in Figure 5. I said earlier that Billy’s throwing his rock did not cause Suzy’s rock to reach the bottle, for Sally’s just keeping her eyes on the circumstances should not affect our initial verdict. However, there is also an intuitive pull toward the opposite claim that Billy’s throwing his rock is indeed a cause of Suzy’s rock reaching the bottle. Certainly, without Billy’s throwing his rock, Suzy’s rock would not have reached the bottle. Notice then how natural we may find it to say that it is in part because Billy threw his rock that Suzy’s rock reached the bottle, and so Billy’s throw is responsible for Suzy’s rock reaching the bottle. Here the guiding intuition is that Billy’s throwing his rock makes a difference to whether Suzy’s rock reaches the bottle. Why doesn’t then Billy’s throwing his rock deserve to count as a cause of Suzy’s rock reaching the bottle?
Some may think it does, and some may not. It seems that our intuitions are not univocal here. One possible explanation would be that the concept of cause is itself not univocal. Indeed, Ned Hall (2004a) makes a plausible case for the thesis that there are in fact two distinct notions of causation, dependence and production, guided by two distinct intuitions. The dependence notion is guided by the intuition that a cause is something that makes a difference to its effects, which is naturally analyzed using counterfactuals. The production notion is guided by the intuition that a cause is something that produces or generates its effects by means of a chain of contiguous events, which is congenial to physical connection approaches that analyze causation as involving the transfer of some (conserved) physical quantities, like energy or momentum (see, e.g., Fair 1979, Salmon 1994, Dowe 2000). Production and dependence are typically co-instantiated—that’s why the distinction between them had previously gone largely unnoticed—but sometimes they are not. In fact, as Hall notes, what cases of preemption and double prevention show are the ways the two concepts of causation can come apart. Preemption cases uniformly exhibit production without dependence. In the case of Billy and Charlie, for example, Billy’s throw produces the shattering, but the shattering does not depend on Billy’s throw. Double prevention cases uniformly exhibit dependence without production. In the case of Suzy, Billy, and Sally, for instance, Billy’s throw does not produce the flight of Suzy’s rock to the bottle, but the flight of Suzy’s rock does depend on Billy’s throw.

I think there is a good deal of plausibility in this explanation and in the distinction between the two concepts of causation. The distinction also gives us a possible strategy for dealing with the overdetermination problem. In addressing the exclusion argument regarding mental causation, Loewer (2007, 255-6) writes:

[T]here is no problem of overdetermination if causation is understood as dependence. On Lewis’s account of counterfactuals a particular event [...] can depend on many co-occurring events. The motions of one’s body, for example, the motions of a person’s arms and hands when reaching into the refrigerator, depend counterfactually both on her mental
states (which snacks she wants) and on her brain (and other bodily) states and on a myriad of other states and events. Also, the kind of “overdetermination” involved in B depending on both M and P is like neither the two assassins kind nor the production kind. In particular there is no temptation to say that if B depends on P it can’t also depend on M since “there is no work for M to do” where, B, M, and P refer to a particular human behavior, its mental cause, and the underlying physical cause, respectively. It is not easy to make this line of argument precise. Two main, related ideas seem to drive this approach, however: First, the kind of overdetermination involved in mental causation can be understood as overdetermination by dependence causes, which Loewer finds unproblematic. Second, the exclusion problem rests on the production conception of causation, which Loewer believes “fits ill with contemporary physics” (2002, 661). Let us consider these ideas in turn.

4.2. Overdetermination by Dependence Causes? As we noted earlier, there are parallels between the difficulties faced by counterfactual analyses of causation and those faced by counterfactual analyses of overdetermination. We also saw that the lesson to be learned from the troubles for the counterfactual approach to causation might be that there are in fact two kinds of causation, production and dependence, rather than that causation is not analyzable in terms of counterfactuals. In a parallel fashion, then, it might be thought that the lesson from the troubles for the counterfactual approach to overdetermination might be that there are in fact two kinds of overdetermination, the production kind and the dependence kind, rather than that overdetermination is not analyzable in terms of counterfactuals.

Consider the case of two sergeants represented in Figure 6, in which two sergeants order the soldiers to advance, so the soldiers move ahead; but if one of the sergeants had not ordered the advance, the major would have ordered a retreat, which would have caused the soldiers to retreat. I said earlier that the two sergeants’ orders overdetermined the soldiers’ advance, for the major’s unactualized intervention should not affect our initial verdict. However, there seems also an intuitive pull toward the claim that there is no
causal redundancy about the sergeants’ orders with respect to bringing about the soldiers’ advancing. Certainly, each of the two sergeants’ orders makes a difference to the soldiers’ advance in the sense that without either one of them, the soldiers’ advance would not have occurred. In other words, both of the sergeant’s orders were necessary to the occurrence of the soldiers’ advance and so, it might well be said, they jointly caused it, not overdetermined it. Yet, on the other hand, the soldiers’ advance really looks overdetermined; there indeed seems to be some sort of causal redundancy in the sergeants’ orders. It might be thought, then, that the right thing to say here, which explains our equivocal intuition, is this: The sergeants’ orders indeed over-determine the soldiers’ advance, but only in the sense of production; in the sense of dependence, they jointly cause it.

If this is right, and there are in fact the two kinds of overdetermination, here is a seemingly plausible response to the overdetermination problem: It would be highly implausible to hold that every case of mental causation involves overdetermination if the overdetermination is thought of as the production kind. This really looks like an ugly picture that involves pervasive causal redundancy in the world. But this is not the kind of situation we have in the case of mental causation. Nonreductive physicalists can hold that mental causation involves multiple dependence causes, and there is nothing objectionable about an effect’s being overdetermined by dependence causes. Our intuition against massive overdetermination stems from considering the situation as one of overdetermination by production causes. However, as Loewer claims, “The kind of overdetermination [nonreductive physicalism] requires is innocuous when causation is understood in terms of counterfactuals” (2002, 661).

An immediate problem with this line of thought, however, is that it is hard to give content to the idea of overdetermination by dependence causes in the first place. In arguing that there is no problem of overdetermination if causation is understood as dependence, Loewer claims that on the dependence notion of cause, an event can be caused by multiple concurrent events. Of course, any typical event will normally have many causes, which any adequate conception of cause should be able to accommodate. Why did the crash happen?
Because the tire was bald, the road was icy, and the driver was drunk. These causes, however, jointly cause the crash—they work together to bring it about. In the case of mental causation, however, the mental cause and the physical cause are not plausibly considered as jointly causing the effect. That is why the overdetermination worry arises, and why Loewer says that some kind of “overdetermination” is involved in mental causation, and argues that the dependency conception of cause “countenances overdetermination” (2002, 659). But how exactly should we understand the kind of overdetermination Loewer thinks is countenanced by the dependence notion of cause? What might instances of overdetermination by dependence causes look like?

Earlier in the chapter I resisted the charge that there might be an incoherence in the idea of overdetermination. I do believe, however, that the idea of overdetermination by dependence causes lacks a coherent sense. The reason, briefly, is this. On the one hand, in order for something to be a cause of an effect, the dependency conception requires that it should make some difference to the effect. On the other hand, in order for it to genuinely overdetermine the effect, it should make no difference to the effect. It seems then that the dependency conception of cause leaves no room for the notion of overdetermination. The very notion of overdetermination by dependence causes looks self-contradictory.

4.3. Does the Overdetermination Argument Presuppose Production Causation?

Pace Loewer, then, the dependency conception of cause does not accommodate overdetermination. Rather, it cannot even give a coherent sense to the idea. There is an interesting

---

9This example is from Lewis (1986b, 214).

10Perhaps this is too quick. For the dependency conception might be messier than the “two concepts” story suggests. Consider, for instance, a case where some process from \( c \) to \( e \) is under two distinct threats, \( t_1 \) and \( t_2 \): Neither threat by itself is sufficient for preventing \( e \); but if neither \( t_1 \) nor \( t_2 \) is blocked, they will jointly prevent \( e \). Suppose \( d_1 \) prevents \( t_1 \), and \( d_2 \) prevents \( t_2 \). There is some intuitive pull to saying that \( d_1 \) and \( d_2 \) both count as dependence causes of \( e \), though \( e \) depends on neither of them; After all, something had to block the threat posed by the joint action of \( t_1 \) and \( t_2 \), and \( d_1 \) and \( d_2 \), between them, did so. If they both count as causes, then it seems that they are overdetermining causes, since the action of each by itself was sufficient to safeguard \( e \). This kind of example, however, does not threaten my crucial claim, that mental causation cases don’t feature dependence overdetermination. Presumably, even if one can make sense of the kind of dependence overdetermination involved here, it isn’t plausible to suppose that standard mental/physical cases feature that kind of overdetermination. Dependence overdetermination is either unintelligible or anyway implausible as a model of mental causation.
dialectical nicety here, however. If the dependency conception renders the idea of overde-
termination unintelligible, doesn’t that mean that the notion of overdetermination itself pre-
supposes the production conception of cause? Or does it mean that the overdetermination
argument rests on the production conception? Doesn’t this all mean that there would sim-
ply be no problem about overdetermination if mental causation is understood as involving
dependence causes?

I believe this line of thought is misguided. For one thing, it is doubtful that the idea of
overdetermination itself presupposes production. There are apparently cases of overdeter-
mination involving omissions. Consider, for example, a case in which your sickness and
your car’s breakdown overdetermine your absence from school. This is a case of overde-
termination of omission—that is, the overdetermined effect is an omission. In this case,
however, there is no productive causal relation between the overdetermining causes and
their effect connected by a spatiotemporally continuous chain of events. Also consider
a case in which your not watering the plant and not providing it with enough sunlight
overdetermine its death. This is a case of overdetermination by omissions—that is, the
overdetermining causes are omissions. In this case, again, there is no productive physical
connection between the overdetermining causes and their effect. This indicates that it is
wrong to think that overdetermination doesn’t make sense unless we think of causation as
production.

More importantly, even if overdetermination requires production, it doesn’t seem quite
right to think that we can just make the overdetermination problem go away by rejecting
the production notion and moving to a dependency conception of cause. To begin, it bears
repeating that the leading idea of the overdetermination argument is very simple. Consider
a case of causation not involving preemption. Suppose in this case two events are thought to
be two distinct causes of an effect, but they are not plausibly thought of as jointly causing
the effect. Then, some sort of causal redundancy appears to be involved, and the worry
about overdetermination arises. It seems that here we are only appealing to a clear and
simple intuition, an intuition involving the distinction between overdetermination and joint
causation. Imagine again a situation in which two sergeants give orders to their soldiers. We seem to have perfectly clear intuitions about the difference between the case in which each order is individually sufficient to move the soldiers and the case in which each is necessary for the soldiers’ advance. And if it looks as though every case of mental causation is like the former, this looks problematic. That is all that is needed to generate the overdetermination worry.

I think that the distinction between overdetermination and joint causation is just part of our commonsense causal thinking. It is something that any account of causation intended to track our ordinary notion of cause had better accommodate and explain. I don’t think, however, that the dependence notion of cause does justice to the distinction. One way of highlighting this is to consider again the case of two sergeants represented in Figure 6, and modify the situation so that both of their orders are in fact necessary to move the soldiers, even if the major isn’t there. Notice that the dependence notion gives the same verdict, that the sergeants’ two orders jointly cause the soldiers’ advance, whether each order is supposed to be sufficient for the soldiers’ advance, as we originally assumed, or merely necessary for it, as we now imagine. It thus does not discriminate between overdetermining causes and joint causes properly; where commonsense finds causal redundancy, the dependence notion sees nothing peculiar.

Note also that the two orders count as joint causes of the soldiers’ advance merely in virtue of jointly preventing the potential preventer of the effect; without the unactualized intervention by the major, the soldiers’ advance depends on neither. So in paradigm examples of overdetermination, like the case of Suzy and Billy, the dependence notion says that neither Suzy’s nor Billy’s throw is a cause of the bottle’s shattering. Observe, furthermore, that this does not change even if we build asymmetry into the example by supposing that Suzy’s rock gets to the bottle first, thus causing its shattering; again, the shattering depends on neither throw. This means that on the dependency conception of causation, there is no difference in causal structure between overdetermination and preemption.
Obviously, the overdetermination problem presupposes the intuitive notion of overdetermination. And maybe that notion carries a commitment to a certain conception of causation. Indeed, some sort of production conception seems to be operative in our judgments about the foregoing examples involving preemption and overdetermination. I’m not suggesting that our operative conception there be understood as involving the transfer of some conserved quantity or any other specific sort of process. That is a substantive claim that would need to be argued for. Rather, we can use the label ‘production conception’ for whatever conception it is that is operative when we judge that overdetermination cases exhibit a different causal structure from cases of preemption, or from joint causation. It may be said, then, that the overdetermination problem really does depend on the production conception of cause.

Does this mean that the problem is resolved when we reject the production conception and move to a dependency conception? In a sense, yes, if we can really reject the production conception altogether, in favor of a dependency conception. However, what foregoing considerations suggest is rather that we cannot plausibly do so. If a dependency conception can neither make sense of overdetermination nor distinguish it from preemption or joint causation, that is a deficiency of that conception. It gives us a prima facie reason to reject the conception. Certainly, we cannot make a problem go away simply by moving to a conception of the notions involved that does not even recognize the problem. The right thing to say seems to be that on the dependency conception, we cannot even sensibly ask whether the effect of a mental cause is overdetermined.

Loewer has claimed that the production conception of causation fits ill with contemporary physics. To be sure, we should not suppose that our ‘folk’ conception of causation will survive utterly unchanged as our best physical theory improves. But it is hard to believe that physics, by itself, can correct our ordinary causal notions involving the difference in causal structure between overdetermination, preemption, and joint causation. It would be quite surprising, for example, if fundamental physics renders problematic our belief that it is Suzy’s throw alone, not Billy’s, that is a cause of the bottle’s shattering, given that it
arrives first. Here I do not want to rule out the possibility that it might turn out that cau-
sation is in fact dependency, and it is rather the very notion of overdetermination that is
problematic. (For example: Why does the dependency conception see nothing where we
see causal redundancy? Because causal redundancy is just in the eye of beholder.) So I
think the following conditional might well be true: There is no problem of overdetermi-
nation if causation just is dependence. But this is a big if. There are many challenges for
explaining causation in terms of counterfactual dependence. Perhaps what I have done is
just to articulate one of them, that the dependency conception does not yield an appropriate
sense of causal redundancy or overdetermination.

5. Closing: A Hybrid Causation View?

I have tried to show why a simple appeal to the dependency conception of cause is not much
help in dissipating our worry about overdetermination, but rather lulls us into ignoring it.
If the dependence notion cannot make sense of (let alone capture) our intuitive idea of
overdetermination, while the production notion is able to do so, we should say that the
dependency conception leaves us unable even to state what the problem is, unless the idea
of overdetermination is itself incoherent.

There is a dialectical point worth noticing. It may have been tempting to think that a
proponent of the overdetermination problem has a burden of showing that the production
conception is the right approach to causation. However, once we see that the dependence
notion does not even make sense of overdetermination, this shifts the burden of proof. The
friends of counterfactuals owe us a reason for believing that the idea of overdetermination
is itself problematic or that the dependency conception is anyway the right approach to
causation.

In closing, I would like to briefly mention a possible picture of mental causation that
might emerge quite naturally at this point, which we may call the ‘hybrid causation’ view.
Given the two distinct notions of causation, it is possible to view mental causation as in-
volving both dependence and production causation, such that the mental cause is a depen-
dence cause of its effect, while the underlying physical cause is a production cause of the
effect. The threat of overdetermination arises, in part, because the mental and physical causes cannot be properly construed as jointly causing the effect—that’s why the effect looks overdetermined. However, it may be a mistake to suppose that two simultaneous causes always overdetermine their effect when they do not jointly cause it. This tacit assumption is wrong, it can be argued, especially when the notion of cause involved is not univocal. The exclusion problem is often put this way: What causal work is left for the mental cause to do if the physical cause does all the work? Questions like this, however, may be wrongheaded if we are dealing with different kinds of causation. And there may be nothing wrong with supposing that an event can have multiple distinct causes without being overdetermined, if the notions of cause involved are different. Mental and physical causes, then, need not causally “compete.” Or so it could be argued.\footnote{It might be noted that the hybrid causation view need not claim that dependency is not involved in underlying physical processes; it may well be that the underlying physical cause is also a dependence cause of the effect. This means that even those, like Loewer, who believe that the production notion is dispensable in describing the basic, fundamental processes of the world may find the hybrid picture acceptable as a rough-and-ready picture of mental causation, though they would try to give a further account, reductive or eliminative, for the production notion of cause.}

I am inclined to believe that, insofar as the overdetermination threat is concerned, the hybrid causation approach indeed is a good enough way of mooting the threat. This is not to say, however, that the approach works as a solution to the mental causation problem. Recall that the exclusion problem takes the form of a dilemma, one between overdetermination and epiphenomenalism. Then, for any alleged way to defuse the overdetermination threat, there always is a danger that it does so only by compromising the causal efficacy of the mental—the more one compromises the efficacy of the mental, the less it sounds like its effect is overdetermined. Note that on the hybrid causation view, the mental does not cause the effect \emph{in the same way} that the physical does. It appears, then, that the view exposes itself to the charge of epiphenomenalism. In general, once you abandon homogeneity of mental causation and physical causation, you risk the charge that your view might be just a disguised form of epiphenomenalism.\footnote{See Crane (1995) for a discussion on the effect of denying homogeneity of mental and physical causation. Crane also argues that denying the homogeneity undermines one of the original motivations for physicalism, for “it’s only insofar as mental states have effects \emph{in the very same sense that physical states have effects} that we need to think of them as physical states” (1995, 235).}
There is another source of epiphenomenalist worry for the hybrid approach. Even if dependency is sufficient for causation, it may not be sufficient for mental causation. It may be that, as Kim (2007) argues, we care about mental causation because we care about human agency, and agency involves a productive conception of causation. If agency requires production, then the hybrid causation view cannot be an adequate account of mental causation, even though it may be a good enough way of saving the efficacy of other higher level, special-science properties.
CHAPTER 4

Mental Causation as Joint Causation

In this chapter, I explore the view that mental and physical events *jointly* cause physical events. This kind of view has usually been taken as a non-starter. It is widely assumed that although the view easily escapes the exclusion problem, it does so only at the high price of violating physical closure. As I show, however, this is doubly mistaken. First, once we get clear on two sources of the exclusion problem, *physical closure* and *supervenience*, we can see that the supervenience problem still arises for the joint causation view. Second, the view is consistent with a moderately strong principle of causal closure that respects much of our physicalist commitments. So I believe there is much to be learned from considering what the joint causation view really amounts to, and exactly what its commitments are. I discuss how we can motivate and make sense of the view within an emergentist picture, and compare it with a prominent nonreductive physicalist conception of mental causation, the overdetermination view. Specifically, I distinguish between the *exclusion* problem and the *supervenience* problem, and argue that the joint causation view fares betters than the overdetermination view vis-à-vis both problems.

1. Causal Exclusion: A Problem for Physicalists or an Argument for Physicalism?

The exclusion problem arises from certain physicalist commitments that many find compelling. Suppose the physical realm is causally closed: every physical thing that happens can be accounted for in purely physical terms. Then how is mental-to-physical causation possible? As it is often put, there seems no room left for the mental to causally contribute to physical happenings, unless it reduces to the physical. Physicalists, then, seem left with a choice between reductionism and epiphenomenalism, unless they are willing to say that every case of mental causation involves causal overdetermination (e.g. Kim 1989, 1998, 2005; see also Malcolm 1968, Peacocke 1979, Schiffer 1987).
Thus it looks as though the exclusion problem is a problem for physicalists. Jaegwon Kim stresses that the exclusion problem arises for anyone with a “broadly physicalist outlook” and in fact “strikes at the very heart of physicalism” (1998, 30). Indeed, the exclusion problem highlights how difficult it is to find a way of accommodating the mental, and its causal efficacy, within a physicalist framework without sacrificing the distinctness of the mental. So Kim has used the problem to push those who want to stay with physicalism toward a form of reductionism, arguing that nonreductive physicalism is not a stable position. The exclusion problem, then, seems to be an internal problem within physicalism.

However, the exclusion problem (or something like it) is also sometimes presented not as a problem but as an argument for physicalism. As Karen Bennett (2008) rightly observes, although the exclusion problem is usually used to defend a version of physicalism (reductive physicalism) against nonreductive physicalism, it has also been used to defend physicalism against anti-physicalism (e.g. Papineau 1995, 2001). A question then arises: How can it be used to argue for physicalism if it is a problem that arises from physicalistic commitments? Whose problem is the exclusion problem?

Like other philosophical problems, the exclusion problem arises from a set of assumptions we accept, either explicitly or tacitly—in this case, ones that are all central to nonreductive physicalism. In particular, causal closure of the physical plays a crucial role in the exclusion argument. Since this principle is commonly taken to be an essential tenet of physicalism, it is clear that any physicalists, including nonreductive physicalists, should care about the exclusion argument. But what about dualists? Where does the principle of physical closure come from, and why should anyone believe it? Perhaps, full-blown dualists might well want to reject physical closure; certainly, for example, Cartesian interactionist dualists would reject it. At least they do not contradict themselves in believing that some physical effects are produced by purely mental causes. It seems then that the exclusion argument has little grip on already committed, anti-physicalist dualists who reject physical closure.
Now, most contemporary dualists are not that blatantly anti-physicalistic. It strikes most of us that physical closure is not something we can easily dismiss, and most present-day property dualists share this concern. Some dualists like David Chalmers (1996, 150) explicitly endorse physical closure and take the exclusion argument seriously. However, what is not clear is whether serious dualists *should* accept physical closure. Consider Chalmers’ “naturalistic” dualists or classical British emergentists. Do they have a compelling reason to embrace physical closure? If this isn’t clear, it is not clear whether the exclusion argument is as forceful as an argument for physicalism (against serious dualism) as it is as an argument for reductive physicalism (against nonreductive physicalism).

So an interesting issue here is whether the exclusion argument exerts the same pressure on a serious form of dualism, like emergentism, toward reductionist physicalism as it does on nonreductive physicalism. Kim (1992, 1993) has argued that nonreductive physicalism and emergentism are in the same trouble over the exclusion problem, due to their shared commitment to physical closure and downward causation. In arguing this, Kim assumes that physical closure is among the basic tenets of emergentism. The same goes for Tim Crane when he holds that “the problems for emergentism—especially the problems of mental causation—are also problems for nonreductive physicalism, and they are problems for the same reason” (2001, 207).

Unlike Kim and Crane, Bennett (2008) leaves open the possibility for emergentists to reject physical closure, and she claims that emergentism and nonreductive physicalism are not in the same boat vis-à-vis the exclusion problem. Her reason for this claim, however, has little to with considerations about physical closure. Her argument is rather that even *if* emergentism and nonreductive physicalism share the commitment to physical closure, they are not in the same boat. Specifically, Bennett argues that nonreductive physicalism has a promising solution to the exclusion problem, a physicalist one, that is not available to emergentists. Thus although emergentism may indeed be in trouble with the exclusion argument, she claims, nonreductive physicalism is not.
I believe, however, that Kim and Crane are right that emergentism and nonreductive physicalism are in the same boat vis-à-vis the exclusion problem if they are both committed to physical closure. Given physical closure, both are indeed in trouble with the exclusion problem. However, I don’t think that they really share the commitment to physical closure. It appears, at least, that emergentists can coherently deny it. So, like Bennett, I believe that emergentism and nonreductive physicalism are not actually in the same position vis-à-vis the exclusion problem.

However, pace Bennett, I don’t think that nonreductive physicalism is better positioned to deal with the exclusion problem than emergentism. I think exactly the opposite is true. As I will argue, the exclusion problem is not really a problem for emergentists, though it is a genuine problem for nonreductive physicalists. This is not to say that the exclusion problem simply does not get off the ground for emergentism. The problem does arise for emergentists—even if they reject physical closure, what I’ll call the supervenience problem still arises for emergentism. But emergentists have a coherent, well-motivated response to the supervenience problem that is not available to nonreductive physicalists. So I believe that although the exclusion argument is indeed a good argument for reduction against nonreductive physicalism, it is not so effective as an argument for physicalism against dualistic emergentism. One primary goal in this chapter is to articulate this point.

The structure of the rest of the chapter is as follows. In section 2, I first describe what I take to be essential to emergentism and nonreductive physicalism, and then motivate the joint causation view of mental causation in an emergentist picture. In section 3, I briefly explain how downward joint causation can be understood in terms of the idea of “configurational” force (McLaughlin 1992). In section 4, I distinguish two versions of physical closure and show that the joint causation view is consistent with a moderately strong version of physical closure. In section 5, I compare the joint causation view with a popular nonreductive physicalist view of mental causation, the overdetermination view, and discuss how they fare differently under the pressure of the exclusion argument. In
section 6, I distinguish the supervenience problem from the exclusion problem and argue that nonreductive physicalism is really in trouble with them, but emergentism is not.

2. Emergentism and Nonreductive Physicalism

Both emergentism and nonreductive physicalism hold a physicalist thesis that the mental supervenes on the physical. And both are anti-reductionist in denying that mental properties are identical, or reducible, to physical properties. They also hold that these irreducible mental properties have causal efficacy. Thus, both positions embrace all of the following:

- **Supervenience:** Mental properties supervene on physical properties
- **Irreducibility:** Mental properties are distinct from physical properties.
- **Efficacy:** Mental properties are causally efficacious.

This means that emergentism might be thought of as a version of nonreductive physicalism, or vice versa. However, most nonreductive physicalists want to differentiate their position from emergentism. Indeed, nonreductive physicalists and emergentists are different in what kind of *Supervenience* they endorse and also in their reasons for accepting *Irreducibility*. This means that they can take different strategies in arguing for *Efficacy*.

First, nonreductive physicalists hold that the mental supervenes with *metaphysical necessity* on the physical. They usually view the mental-physical relation as realization, constitution, or some determinate-determinable relation, so that mental properties are in some very strong sense determined by physical properties. (Though they usually buy into the familiar multiple realization argument against reduction.) So when nonreductive physicalists endorse *Irreducibility*, they only mean that mental properties are not identical to physical ones, or that mental and physical properties are numerically distinct. Thus, nonreductive physicalism is a very attenuated form of dualism.

In contrast, emergentists believe that mental properties, as emergent properties, are novel, extra additions to the world’s furniture. On emergentism, when certain physical states give rise to certain mental states, they do so by virtue of instantiating a fundamental
law of nature. Mental facts, then, are conceived as something “over and above” physical facts. So when emergentists endorse *Irreducibility*, they mean that mental and physical properties are two irreducible basic kinds of properties, though they are lawfully connected by bridge laws or trans-ordinal laws. Here bridge laws are supposed to be brute and ultimate facts to be discovered empirically. Emergentists hold that the mental supervenes on the physical with *nomological* necessity, not with metaphysical necessity. (I will use ‘emergentism’ as a label for the position characterized here.)

There are two related worries here. One is about whether metaphysical supervenience is compatible with *Irreducibility*. That is, once one accepts metaphysical supervenience, this raises doubt on whether one is indeed a “nonreductive” physicalist. The other is about whether nomological supervenience is enough to vindicate our general physicalistic intuitions. That is, once one denies metaphysical supervenience, it raises doubt as to whether one deserves the label “physicalist”—this is why emergentism is not usually considered a form of physicalism. Here I will simply assume that metaphysical supervenience does not itself amount to reduction. My concern is rather with the question of whether nomological supervenience, together with other plausible assumptions, is sufficient to capture our physicalist commitments.

In saying ‘other plausible assumptions,’ I have in mind primarily the following thesis:

*Physical Closure*: The physical domain is causally closed.

This says that the physical realm is a causally self-sufficient system. However it is to be spelled out, any good physicalist would accept the thesis. Once one is committed to something like *Physical Closure*, however, it invites the challenge: If physical events have complete physical causal histories, how could there be mental causation?

In response to this challenge, many nonreductive physicalists have attempted to save mental causation by appealing to a tight metaphysical relation between the mental and the physical. It is often argued that mental and physical causes are so intimately related that
they can both be causally sufficient for the same effect without overdetermining it (e.g. Yablo 1992, Shoemaker 2001, Pereboom 2002, Bennett 2003). On the nonreductive physicalist conception of the mental, mental causes are realized by, constituted by, or just metaphysically supervene on physical causes. Then, the argument goes, mental and physical causes are not distinct or independent in the way that two causes in standard overdetermination cases are (e.g., in firing squad cases in which more than one bullet hit the victim’s heart at the same time). Mental causation always involves more than one sufficient cause, but there is no causal competition or overdetermination.

An alternative way of pushing the same line is to grant that mental causation indeed involves a kind of overdetermination, but deny that this is problematic. It is of course incredible, the thought goes, to think that every case of mental causation is a case of overdetermination if it is thought of as like a firing squad case. But this is not the sort of overdetermination involved in mental causation. As already noted, the mental cause is metaphysically tightly connected with the physical cause, and this sort of overdetermination is indeed widespread and unobjectionable (e.g. Mellor 1995, Loewer 2002, Schaffer 2003, Sider 2003). This strategy, unlike the one described above, explicitly acknowledges that mental causation always involves overdetermination. However, the difference between them seems more terminological than substantive. The important thing is that both strategies commonly, and crucially, turn on the idea that (i) the mental/physical case is importantly different from the standard examples of overdetermination in that the relation between the mental and physical causes is much more intimate than the relation between, say, two bullets in the firing squad, and that (ii) this difference is what makes it perfectly acceptable to claim that mental causation always involves more than one sufficient cause. Let us call this line of approach the overdetermination approach to mental causation.

I find this line less than fully satisfying. It isn’t clear to me why we should think that the tight relation between the mental and the physical defuses the problem, or why it makes the kind of overdetermination involved less troublesome. It is not my aim here, however, to directly assess this nonreductive physicalist view. Instead, I want to draw attention to
an emergentist view of mental causation that has been almost entirely neglected, the joint causation view, and compare it with the overdetermination view, drawing some lessons about the exclusion problem.

Emergentism holds that emergent properties are new, net additions to the world. If one believes in such properties, it seems entirely natural to believe that they have their own causal powers—in fact, it was Samuel Alexander (1920), a leading British emergentist, who equated the reality of an entity with its having causal powers. If they don’t introduce new causal structures into the world, what could be the point of these extra entities? C. Lloyd Morgan, another leading early emergentist, writes:

> When some new kind of relatedness is supervenient (say at the level of life), the way in which the physical events which are involved run their course is different in virtue of its presence—different from what it would have been if life had been absent. (1923, 16)

Morgan’s commitment to emergents’ novel causal powers is explicit: Once higher-level properties have emerged, they make a difference to ongoing physical events at the lower-level, so that without these higher-level properties, the physical events would have run their course differently.

I believe that emergentists are willing to endorse, and perhaps are committed to, the claim that emergents’ causal powers go beyond the powers of their emergence bases. And I think the most natural and straightforward way of understanding the claim is this: Once an emergent arises from its underlying physical base, the emergent and its physical base jointly cause their effect. Suppose a mental property, M1, emerges from its physical base, P1, and they are followed by the subsequent physical event P2. The idea is that M1 and P1 jointly cause P2 so that if P1 were to occur without M1, P2 would not have occurred. It is clear then that emergents make causal contributions over and beyond what their underlying bases contribute.¹
The joint causation view holds that when the mental causes something, the mental and its underlying physical base jointly cause the effect. On this view, it is not true that the physical underlying base, absent the mental cause, is causally sufficient to bring about the effect. Physicalists, then, who insist on Physical Closure might find this view unacceptable from the outset. They might think of the view as submission to the exclusion problem rather than resolution of it. After all, the main point of the exclusion argument is that we cannot have both causal efficacy of the mental and causal closure of the physical. The joint causation view may sound like just a recommendation to choose the former at the cost of the latter.

Indeed, I am prepared to agree that irreducible causal efficacy of the mental is not compatible with Physical Closure. This is why I believe that the exclusion argument is a good argument for reduction against nonreductive physicalism. However, emergentism fares differently vis-à-vis the exclusion argument. What I want to argue here is that the joint causation view is not a mere ad hoc maneuver to avoid the exclusion problem on pain of violating Physical Closure. Rather, it is a well-motivated emergentist solution to the exclusion problem that is not available to nonreductive physicalists. The joint causation view does not solve the problem on the cheap, by simply denying Physical Closure. As we will see, it is consistent with a version of Physical Closure that saves much of our physicalist intuitions. Moreover, the supervenience problem still applies to the joint causation view, though I will eventually argue that the joint causation view has a good answer to the problem, while the overdetermination view doesn’t.

3. Making Sense of Downward Joint Causation

On the joint causation view, emergent properties causally influence lower-level physical processes as partial causes, so that physical events sometimes need help from higher-level causes to bring about their effects. One might ask how this is possible. Does that mean that

1In the quote above, Morgan is ambiguous between two readings: (i) if P1 were to occur without M1, P2 would not have occurred; (ii) if M1 had not occurred, P2 would not have occurred. The second reading is not committed to the joint causation view.
the mental fills in causal gaps in physical causal processes? By what mechanism could that happen?2

In his detailed and comprehensive discussion of British emergentism, Brian McLaughlin (1992) provides an account of how the emergentist idea of downward causation can be explained in terms of “configurational” force. According to McLaughlin, emergent properties could be basic force-generating properties, and emergent causal powers could be grounded in configurational forces, “fundamental forces that can be exerted only by certain types of configurations of particles, and not by any types of pairs of particles” (1992, 52). Configurational forces are, therefore, unlike the gravitational force that holds between any two objects. They also contrast with forces exerted only by (certain specific kinds of) basic particles, like the nuclear force or electromagnetic force. Such forces as frictional force and Van der Waals forces also cannot be regarded as configurational forces because they are not fundamental; they are all derivable from the electromagnetic force. So configurational forces are supposed to be additional, fundamental forces that can only be exemplified by a limited range of matter with a certain type of (complex) structure.

What might such configurational forces be like? Here is an example: Imagine a simple world that consists of a single type of basic particle which is governed by the gravitational force. Each particle accelerates other particles with the force that is proportional to the product of the two masses and inversely proportional to the square of the distance between them. In addition, suppose that it is a law of nature in this world that when particles aggregate by gravitation into a certain critical amount of mass, say, 1kg, the aggregate exhibits a novel property, say, being yellow. And once an aggregate becomes yellow, it generates a novel fundamental force and influences movements of other particles. Suppose, for example, when an aggregate becomes yellow, it exerts, say, the yellow-force on every other mass that is proportional to the distance between them.

2For example, Eric Marcus (2005, 42) writes: “By thinking of mental causes as picking up the causal slack left by physical causes, we adopt what we may call the telekinetic view of mental causation: i.e., mental causation as some kind of spiritual pushing. We stand to our bodies, or perhaps just to our brains, in the same relation that Yuri Geller purports to stand to his fabulous bending spoons.”
There seems nothing inherently mysterious or incoherent with this sort of emergent configurational force. Indeed, as McLaughlin (1992) points out, configurational forces can be easily accommodated within the framework of classical mechanics. In classical mechanics, the general laws of motion do not place any limit on what kind of forces can be exerted on an object; a configurational force, which specifies one basic factor that influences the acceleration of an object, could be a component force-vector of the net force on the object. And the classical framework allows the idea that there may be basic force-generating properties that are possessed only by a limited range of kinds of matter—for example, Coulomb’s law applies only to objects with charges.\textsuperscript{3} If so, in order to determine the net force on an object around a yellow aggregate in the imagined world, and thereby to determine the resultant acceleration of the object, we should take into account not only the gravitational force but also the yellow-force exerted on the object. That is, to explain the acceleration of the object, we would have to invoke not only the aggregate’s mass but also its being yellow.\textsuperscript{4}

This illustrates how an emergent property and its underlying physical property could jointly cause an effect; the aggregate’s mass and being yellow jointly cause the acceleration of the object. In this picture, emergents make their genuine causal contributions to the lower-level processes by generating fundamental causal forces, “configurational” forces, over and above physical causal forces. Let us assume that McLaughlin is right that the notion of configurational force is itself perfectly coherent—he believes that the fall of British emergentism is due to empirical, rather than a priori philosophical, reasons. Even if the idea of mental causation as downward joint causation is coherent, however, the resulting world view might not be palatable. It may look too anti-physicalistic to accept.

\textsuperscript{3}Similarly, as McLaughlin also points out, configurational forces are compatible with quantum mechanics as well; the Hamiltonian in Schrodinger’s equation implies no constraint on the kind of forces (or energies) of the physical system. McLaughlin argues further that they are compatible with relativity theory in that “configurational forces need not involve any violation of [the conservation of mass-energy] . . . Configurational forces could involve various compensating shifts in mass and energy that maintained conformance to the principle of mass-energy” (1992, 74).

\textsuperscript{4}One might ask why the situation cannot be described without positing special configurational forces, by modifying or supplementing the law of gravitation. I will later return to this question.
4. Strong Closure and Moderate Closure

On the joint causation view, sometimes a mental event is a necessary constituent of a full cause of a physical event. The view then seems to conflict with Physical Closure; a complete causal explanation of the physical effect would, at least partially, go outside the physical domain. But there are a variety of nonequivalent formulations of Physical Closure, and we need to be careful about which ones the emergentist picture violates and which ones not.

The basic idea behind Physical Closure is that every physical effect is fully determined or necessitated by its physical antecedents. Let us first consider perhaps the simplest formulation of this idea. It says that every physical effect has a physical cause. This would be too weak to capture the intended physicalist intuition. It allows a causal chain consisting of a physical effect caused by a mental event which in turn is caused by a physical event. So Physical Closure is often put like this: If a physical event has a cause that occurs at $t$, it has a sufficient physical cause at $t$ (e.g. Kim 2006, 195). This says that every physical event (that has a cause at all) has a sufficient physical cause (at every time at which it has a cause). There will be a number of ways to state this principle (in particular, the phrase ‘sufficient’ physical cause) more precisely. One way of stating the principle is this:

*Completeness:* All physical effects are fully determined by law by prior physical occurrences.

This is the principle Papineau (2001, 8) calls “the completeness of physics.”5 The idea is that fixing prior physical conditions of a physical event guarantees the occurrence, or fixes the chance, of the physical event. If Completeness is true, then all physical effects are due to physical causes, and this appears to contradict the joint causation view. However, there is a construal of the principle that renders it consistent with downward joint causation.

---

5Papineau also sometimes calls it the “causal closure” of physics or of the physical (e.g. Papineau 2008). Its indeterministic version runs that “the chances of all physical occurrences are fully determined by prior physical occurrences” (2001, 33, n.2).
Suppose our world is indeed as the emergentist conceives it. The world is governed not just by physical laws but also by fundamental psychological laws, purely psychological or psychophysical, and there indeed occurs downward joint causation. Consider again the situation in which M1 emerges from P1 and they jointly cause P2, which can be depicted by the following figure.

![Figure 1](image)

M1 and P1 jointly cause P2, and the link between M1 and P1 represents a nomological necessitation relation between them.

Here we may suppose that there are nomological necessitation relations underlying the causal relations between P1 and P2 on the one hand and M1 and P2 on the other. It seems then that the occurrence of P2 is fully determined by a prior physical state, namely, P1 in the sense that there is no nomologically possible world where P1 occurs but P2 does not occur. This is because every nomologically possible world where P1 occurs is a world where M1 also occurs, and therefore P2 would occur in such a world. Given the occurrence of P1, the occurrence of P2 seems to be fully determined as a matter of nomological necessity. There is a sense, then, in which downward joint causation meets Completeness; P2 is fully determined by law by P1.

There is of course a reading of Completeness that renders it incompatible with downward joint causation. Completeness says that every physical occurrence is fully determined by law by antecedent physical conditions. What sort of laws are involved here? We can distinguish two versions of Completeness depending on whether it involves merely the laws of physics or the laws of nature:
Strong Closure: Every physical effect is fully determined by laws of physics by prior physical conditions.

Moderate Closure: Every physical effect is fully determined by laws of nature by prior physical conditions.

To see what these amount to, it will be useful to compare these with a weaker version of causal closure:

Natural Closure: Every physical effect is fully determined by laws of nature by prior natural conditions.

Natural Closure rules out the possibility of any super-natural or transcendental causes of physical events. Let us set aside the question of what ‘natural’ means, and assume that, for simplicity, the mental and the physical exhaust the domain of the natural. Natural Closure then says that every physical effect is fully determined by prior physical and mental conditions. It is clear that the joint causation view respects this principle—for a given P2, its occurrence is entailed by the prior natural condition, M1+P1, and the laws of nature. One might think that Natural Closure is the very kind of causal closure that emergentists have reason to accept or are committed to. Indeed, serious property dualists might well be content with Natural Closure; if the mental is a brute, ultimate feature of the world, then it seems that it should figure in our causal explanation of natural events. But I think the emergentist can go further.

Natural Closure allows the possibility in which there may be some physical effects that are not nomologically fully determined by prior physical conditions alone—you will need to invoke some further mental conditions as well to ensure their occurrences. It is this kind of possibility that Moderate Closure rules out. Moderate Closure says that every physical occurrence is entailed by prior physical conditions and the laws of nature. And what we have seen above is that downward joint causation does not violate this principle. As we saw, P1 in the emergentist world nomologically entails M1 and thereby nomologically entails P2 as well. In contrast, downward joint causation does violate Strong Closure. P1 and the laws of physics do not entail P2 by themselves. There would be some physically possible
worlds, worlds that perfectly conform to the laws of physics of the emergentist world, in which P1 occurs but P2 does not—only the psychophysical law linking P1 and M1 (or M1 and P2) is violated in such worlds.

Thus, Strong Closure does not hold in the emergentist world. A complete causal explanation of P2 should include reference to M1, and in this sense physics is not complete in that world. However, given the nomological link between M1 and P1, if you have P1, it is guaranteed that you will have P2—whenever P1 occurs, you thereby have M1, and so also have P2. Thus you can always infer or predict P2 from P1, and there is a sense in which P2 is fully determined by P1 as a matter of nomological necessity. This has an interesting implication.

Imagine a physicist or neuroscientist who tries to explain P2 in physical terms. He might cite P1 in explaining P2 and might (reasonably) believe that there can be a complete explanation of P2 in purely physical terms. He might be totally unaware of P1’s correlate, M1, or believe that M1 is just epiphenomenal. If he believes the completeness of physics or is already committed to something like Strong Closure, he would not care about emergents like M1 in constructing his theories about P1 and P2. Indeed, there would be no direct empirical way of detecting M1’s causal contribution to P2—we cannot set up an experiment showing what would happen if P1 occurs without M1, since it is just (nomologically) impossible. However, if the situation is really as the joint causation view depicts it, the physicist would be mistaken in thinking that P2 has a complete physical causal explanation, though the causal role of M1 would be “invisible” to him to use an expression from E.J. Lowe.6

6In his (2000, 2003), Lowe gives an emergentist, interactionist model of mental causation, and argues that if interactive dualism is true, it would be entirely reasonable to expect mental causation to be “invisible,” in the sense that no “gaps” would be apparent in the causal relations between physical events, and all physical events would seem to have wholly physical causal explanations. As an emergentist model, Lowe’s model of mental causation bears a certain similarity with mine. Like Searle (1992), however, Lowe appeals to a problematic conception of the mind-body relation, that simultaneous causation is the relation between the mental and the physical.
What this illustrates is that neither the issue of whether *Strong Closure* is true nor the issue of whether the mental is efficacious or epiphenomenal is something that can be settled on entirely empirical, scientific grounds. Consider again the yellow-force world. I described the world as a world in which when particles aggregate into a certain amount of mass, the aggregate becomes yellow and exerts the yellow-force. However, one might try to describe the situation without positing special configural forces, by modifying or supplementing the law of gravitation, to provide for the gravitational force generated by aggregates of particles. Being yellow would then appear epiphenomenal and just a “nomological dangler.” But this would be a mistaken description if the world is really as the joint causation view describes it. However, the issue here may not be a purely empirical one. J.J.C. Smart (1959, 155) once aptly pointed out that while the issue between the brain-process theory and, say, the heart-process theory is a purely empirical, scientific one, the issue between the brain-process theory and epiphenomenalism is a philosophical one. I believe the same goes for the choice between the joint causation view and epiphenomenalism in the imagined world and also for the choice between *Strong Closure* and *Moderate Closure*.

Now, if emergentism is not compatible with *Strong Closure*, what should we say about this? I have three points to make. First, I think any serious emergentist should be willing to accept the consequence that downward joint causation is inconsistent with *Strong Closure*. Emergentists hold that in addition to physical laws there are other ultimate laws that link physical properties to emergent properties, so that emergent properties would figure in the network of the ultimate laws of the world. It seems then that emergentism arguably implies the denial of *Strong Closure*. Or, at least, *Strong Closure* goes against the spirit of emergentism.

However, second, in virtue of supervenience of the emergent on the physical, emergentism accommodates *Moderate Closure*. *Moderate Closure* says that every event is fully determined by prior physical events. Admittedly, P1 *does* need the help of M1 to bring about P2. But given the nomological link between P1 and M1, P1 is itself nomologically
sufficient for M1, so there is a sense in which P1 is indeed sufficient for P2—given P1, P2 is nomologically guaranteed. Without supervenience, property dualists would have to settle for *Natural Closure*. With supervenience, however, there is a sense in which on emergentism, physical facts determine all the facts of the world—there are no nomologically possible worlds in which all the physical facts are identical to our world but psychological facts are different from ours. One may still feel that this emergentist picture is not robustly physicalistic. Then so be it. I myself think that naturalism is a more preferable label for emergentism (e.g. Chalmers 1996). In any case, if we get clear on its commitments, the label would not matter much.

5. The Joint Causation View and the Overdetermination View

As we noted earlier, the exclusion argument is sometimes used to argue for reductionism against nonreductive physicalism and sometimes for physicalism against anti-physicalism. We are now in a position to see exactly how emergentism and nonreductive physicalism fare differently vis-à-vis the exclusion argument. One way to put the exclusion argument is that the following claims, each of which is plausible taken on its own, seem inconsistent taken together:

*Irreducibility:* Mental events are distinct from physical events.

*Efficacy:* Mental events cause physical events.

*Physical Closure:* Every physical effect has a sufficient physical cause.

*No Overdetermination:* Physical events are not pervasively overdetermined.

It looks as though we have to give up one of these. Nonreductive physicalists cannot reject *Irreducibility* or *Physical Closure*. They are then left with a dilemma between discarding either *Efficacy* or *No Overdetermination*. Now *Efficacy* is also central to nonreductive physicalism; as Fodor (1990, 156) put it, losing it would be “the end of the world.” It seems then that nonreductive physicalism has no choice but to accept overdetermination.
So the overdetermination view holds that mental causation indeed involves overdetermination, but insists that the sort of overdetermination involved is not problematic because the mental and physical causes are metaphysically tightly related.

The bottom line is this. The overdetermination view endorses Irreducibility, Efficacy, and Physical Closure on pain of abandoning No Overdetermination. They soothe the worry about widespread overdetermination by appealing to the intimate metaphysical relation between the mental and the physical, namely Supervenience.

Unlike nonreductive physicalists, emergentists can avoid the exclusion problem by rejecting Physical Closure. But many emergentists want to also have Physical Closure (e.g. Chalmers 1996). Once they embrace it, however, they are faced with the same dilemma between Efficacy and No Overdetermination as nonreductive physicalists. This is Kim’s and Crane’s point when they claim that emergentism and nonreductive physicalism are in the same position with respect to the exclusion problem. And that is why Chalmers (1996) takes the epiphenomenalists way out regarding qualia (also see Jackson 1982).

However, our discussion shows that emergentists need not, and perhaps should not, be epiphenomenalists without other independent reasons. Given the distinction between Strong Closure and Moderate Closure, there is no compelling motivation for emergentists to embrace Strong Closure, which rather goes against the spirit of emergentism. Thus emergentists are well-positioned to resolve the problem by rejecting Strong Closure and accepting the joint causation view. One might think that rejecting Strong Closure is too high a price to pay. Indeed, it may deprive one of the right to be called a physicalist. But it is not as pricy as it may first appear. Emergentism can still have Moderate Closure in virtue of the tight metaphysical relation between the mental and physical causes.

The bottom line is this. The joint causation view endorses Irreducibility, Efficacy, and No Overdetermination on pain of abandoning Physical Closure. They soothe the worry about the violation of Physical Closure by appealing to the intimate metaphysical relation between the mental and the physical, namely Supervenience.
Given the apparent parallel between the two balance sheets for the overdetermination view and the joint causation view, it may look as though emergentism and nonreductive physicalism are on even ground with the exclusion problem. But they are not. That the four premises of the problem are seemingly inconsistent taken together is really a problem for nonreductive physicalism, but not for emergentism after all. Nonreductive physicalists cannot easily discard any of the premises. But emergentists may justly reject *Physical Closure*. This means that to use the exclusion problem as an argument against emergentism, one had better give an independent, non-question begging reason to accept *Physical Closure*.

In fact Papineau, who takes the exclusion-style argument to be the principal argument for physicalism—he calls it “the causal argument for physicalism”—thinks that there is such a reason. He thinks that *Physical Closure* is not just a physicalist assumption, but something we all should respect, if we are to base our beliefs on our best physical theory. In examining the history of *Physical Closure*, Papineau claims that it is an empirically established thesis. He says that it is a “highly empirical claim, whose acceptance derives from detailed empirical evidence about the causes of physical effects” (2008, 60). If we conceive *Physical Closure* this way, the exclusion problem, or any other argument that invokes *Physical Closure*, should be of concern not just for physicalists but for anyone who want to be scientifically responsible.

However, I don’t think that it is plausible to regard *Physical Closure* as an empirical thesis that “if true, must somehow follow from the findings of science” (Papineau 2008, 55). What we have seen in the previous section is that the issue of whether to accept *Strong Closure* or *Moderate Closure* isn’t settled on purely empirical grounds. Admittedly, it seems true that we have come to believe something like *Physical Closure* from the findings of science. In his discussion of British emergentism, McLaughlin (1992) persuasively argues that the emergentists’ accounts of chemical and biological phenomena were largely rejected for empirical, rather than philosophical, reasons. It may well be true that advances of science at the time—especially, the quantum mechanical explanation of chemical bonding—made
the doctrines of emergentism, particularly those involving configurational forces, redundant and implausible, and led to its downfall. And the emergentists were wrong in thinking that there are chemical or biological configurational forces. Even in a case like this, however, broadly *a priori* and methodological considerations played a larger role in theory choice. It doesn’t seem to me that the truth of *Physical Closure* can be resolved by sheerly empirical findings. I think the principle is a metaphysical thesis, whose plausibility derives from broadly metaphysical considerations, or a methodological, regulative principle that guides practicing scientists. The upshot, then, is this: the exclusion argument is not as forceful as an “empirically based argument for physicalism” against emergentism (Papineau 2001) as it is as an argument for reductionism against nonreductive physicalism.

In this section, we have observed that the tight metaphysical relation between the mental and the physical, *Supervenience*, is crucial for both emergentists and nonreductive physicalists to handle the exclusion problem. The joint causation view chooses *Efficacy, Irreducibility*, and *No Overdetermination* at the cost of *Physical Closure*, whereas the overdetermination view chooses *Efficacy, Irreducibility*, and *Physical Closure* at the cost of *No Overdetermination*. And both appeal to *Supervenience* to soothe the worries about their costs. But it might turn out that *Supervenience* is not the sort of thing they can freely appeal to. Indeed, as Kim (1998, 41) stresses, *Supervenience* is a source of the problem, far from being part of the solution, as hoped for by many nonreductive physicalists. And this is where I think emergentists are in a better position than nonreductive physicalists in a particularly interesting way.

### 6. The Overdetermination Problem and the Supervenience Problem

A brief way of putting the exclusion problem goes like this: Given causal closure of the physical, there seems no causal work left for mental causes to contribute; the underlying supervenience base appears to do all the causal work. A point whose significance has gone unappreciated is that the epiphenomenalist worry here has two different sources. The first is *Physical Closure*. Once we suppose that the physical is causally closed, this seems to imply that there is no causal work left for the non-physical. The second is *Supervenience*. 
Once we suppose that a certain thing’s being itself is necessitated in some robust sense by another thing, then for whatever the former caused, we seem to be able to say, in effect, that the latter caused it.

It is important to recognize that the two sources are independent of each other. On the one hand, once one accepts Physical Closure, one is faced with the epiphenomenalist worry whether or not one accepts Supervenience. Indeed, the exclusion problem is often formulated without Supervenience. As we have seen in the previous section, Supervenience was not among the four premises of the exclusion problem; rather, it was appealed to as a cure for the problem.

On the other hand, once one accepts Supervenience, one is faced with another sort of epiphenomenalist worry whether or not one accepts Physical Closure. In fact, cases of downward joint causation nicely illustrate this. Consider again the case of Figure 1 in which it is supposed that P1 and M1 jointly cause P2. In this case, the exclusion problem—that there is no work left for M1 if P1 does all the work needed—does not arise, since the joint causation view does not suppose that P1 does all the work needed.

However, the threat is still alive that P1 may preempt M1’s causal status for any alleged effect as long as it is supposed that M1 is necessitated by P1. It is true, we are supposing, that P1 needs M1’s help in order to bring about P2. But M1 itself is necessitated by P1. To put it another way, if you have P1, that automatically gives you M1—that is how the joint causation view meets Moderate Closure. There is a sense, then, in which to bring about P2, all you need is P1. The emergentist then seems to face a critical question: If M1 is necessitated or determined by its basal condition P1, why cannot P1 displace M1 as a cause of any putative effect of M1? Why cannot P1 do all the work in explaining why P2 occurred?

Here the source of the worry is that, as Kim (1998, 53) puts it, “the causal status of the dependent event is threatened by the event on which it depends.” So, even if we suppose that the supervenient property and the physical property jointly cause the effect, the threat still remains real, as long as the mental cause has the underlying subvenient, whether the
underlying domain is causally closed or not. Then, the emergentist claim of joint mental causation might turn out to be empty. We may call this problem the “supervenience” problem, distinguishing it from the problem that arises from Physical Closure, which we may call the “overdetermination” problem. The two problems differ in their sources of epiphenomenalist worry. They are two independent problems.\footnote{Kim’s standard way of presenting the exclusion (or supervenience) argument obscures this distinction by mixing up the two problems (e.g. Kim 1998, 2005). Kim comes close to making the distinction when he distinguishes between the “exclusion argument” and the “supervenience argument” in his (2011, 214-220). But he rather puts the distinction in terms of their purposes: The supervenience argument purports to show that mental-to-mental causation presupposes mental-to-physical causation, while the exclusion argument purports to show that mental-to-physical causation is problematic. My distinction here does not concern mental-to-mental causation at all. It is just about the source of the problem: The supervenience problem uses only Supervenience, not Physical Closure, as a premise, while the overdetermination problem uses only Physical Closure, not Supervenience, as a premise. This roughly corresponds to Kim’s two ways of completing the second stage of the supervenience argument in his (2005, 41-45). To fix the terminology, we may use the ‘exclusion argument’ as a generic term for any similar arguments purporting to exclude mental causation.}

What we have seen is that Supervenience might not be the sort of thing emergentists can happily appeal to in handling the overdetermination problem; it might make them fall victim to the supervenience problem. It should be noted that if the supervenience problem is a problem for the joint causation view, it is equally a problem for the overdetermination view, as long as it endorses Supervenience. The heart of the problem is that there is a “tension between vertical determination and horizontal causation”—namely, the former’s threat to void the latter (Kim 2005, 38). That is why the threat remains even if we suppose that M1 and P1 jointly cause P2 (whether the physical domain is closed or not). Then the threat must remain as long as M1 is alleged to be a cause of P2, whether it is supposed to be a joint cause or an overdetermining cause—maybe a fortiori if it is supposed to be an overdetermining cause.

I believe that emergentists are much better positioned here than nonreductive physicalists. Let us first consider the nonreductive physicalist picture. On nonreductive physicalism, M1 is necessitated by P1 as a matter of metaphysical necessity. Given such a strong vertical necessitation relation, once you have P1, you already have M1; there is no possible world in which P1 obtains but M1 does not. To borrow Kripke’s metaphor, once God creates P1, he doesn’t have to do anything further to create M1. If so, all one needs to bring
about P2 is just P1. What significance, then, can we attach to the claim that M1 is a distinct cause of P2? Why not just say that there is but one genuine causal process from P1 to P2, and that our talk of downward causation from M1 to P2 is just another way of representing that process in terms of higher-level mental *predicates*? Without an answer to this question, the causal efficacy of M1 may look simply stipulative.

The emergentist picture is quite different. When emergentists say that M1 is necessitated by P1, it only means that they are nomologically connected. M1 and P1 always concur by the laws of nature, but this does not mean that once you have P1, you already thereby have M1. There are possible worlds in which P1 obtains but M1 not, so God needs further work to create M1 after creating P1. It seems then that there is little conceptual difficulty in supposing that M1 is a distinct cause of P2. More generally, there seems to be little conceptual difficulty in positing both vertical nomological determination and horizontal causation. Indeed, the emergentist conception of downward joint causation purports to illustrate how we can coherently have them both, by supposing that M1 and P1 jointly cause P2. We can, then, put the upshot this way. The stronger the vertical determination relation one has, the stronger the preemption threat is, and so the harder the supervenience problem becomes.

7. Conclusion

So far I have tried to motivate and make sense of an emergentist view of mental causation, the joint causation view, and compare it with the nonreductive physicalist view of mental causation. My conclusion is that emergentism is in a better position than nonreductive physicalism with respect to both the overdetermination problem and the supervenience problem. It seems to me that the joint causation view is metaphysically more intelligible and defensible than the overdetermination view. At least I think the joint causation view should be considered more seriously as an available option than it generally is.
Bibliography


