The London School of Economics and Political Science

# An Explanatory Account of Practical Reasons

Deren Cem Halil Olgun

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics and Political Science for the degree of Doctor of Philosophy, December 2017.

## Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).



I declare that my thesis contains 91,596 words (excluding bibliography and appendices).

# Abstract

If I take my umbrella, having seen that it's raining outside, we might say that my reason for taking my umbrella was that it was raining. However, if I'd believed that it was raining when it wasn't, we might say that my reason for taking my umbrella was that I believed that it was raining. In the first case, my reason for acting seems to be a feature of the world, whilst in the second it seems to be a feature of my psychology.

According to most theories of reasons, we are mistaken about what my reason for acting was in one of these cases. However, I argue, these theories all entail several awkward claims.

I argue that there is a theory of reasons that can reconcile these two accounts of what my reason for acting was without entailing such awkward claims. I argue that what the fact that it is raining and the fact that I believe that it is raining have in common is that, in their respective cases, they each explain why it was rational for me to take my umbrella and why I took it. More generally, I argue that there is at least a sense in which all practical reasons explain why it is, in some respect, rational for the agent to do the actions for which they are reasons.

The major challenge for this account is the claim that only features of an agent's psychology can explain why they act or why it is rational for them to act. I provide a formal construal of this challenge and argue that the fact that it is raining can explain why I take my umbrella and why it was rational for me to do so, by explaining the fact that I believed that it was raining.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

This thesis exists, in large part, because of the advice and support of my friends and colleagues. I am grateful to all of them, but I would like to thank the following people, in particular.

First, I would like to thank my primary supervisor, Christian List. Were it not for his rigorous and precise feedback, this work would have been far poorer, and were it not for his open-mindedness it would have certainly lacked whatever it now has by way of originality. I would also like to thank my secondary supervisor, Richard Bradley, for his astute criticisms, for his encouragement, and for being so generous with his time. And I would like to thank my examiners, Maria Alvarez and Jonathan Dancy, for their thoughtful engagement with this thesis, and for their words of support. Having spent so much time thinking about their work, it was an honour to hear their thoughts on mine.

I am enormously grateful to Bryan Roberts for helping me through several iterations of The Explanatory Exclusion Problem. Without his guidance my elaboration of that problem would have been twice as long and half as good. I am also grateful to Pete Faulconbridge and James Nguyen for their thoughtful responses to an early draft of chapters (XV) and (XIV); their insights transformed my arguments. And I am grateful to Todd Karhu for his considered engagement with all of the myriad problems I put to him; his influence on this thesis is subtle but pervasive.

I would like to thank the donor who funded my scholarship, and who wished to remain anonymous; their generosity made this work possible. I would also like to thank all of my colleagues at the Office of Rail and Road, and Chris Hemsley in particular, for being so supportive and so accommodating.

I am lucky to have enjoyed the friendship of my fellow PhD students. It would be remiss of me not to mention, in particular, and in addition to those already mentioned, Susanne Burri, Goreti Faria, Johannes Himmelreich, Silvia Milano, Mantas Radzvilas and Nicolas Wüthrich. Their companionship through the peaks and troughs of this work was invaluable.

Finally, I thank my partner, Alex Bobocica, who has been supportive, patient, and considerate throughout. I don't know how I would have done this without her.

# Introduction

If I take my umbrella, having seen that it's raining outside, we might say that my reason for taking my umbrella was that it was raining. However, if it hadn't been raining, but I'd mistakenly believed that it had, we might instead say that my reason for taking my umbrella was that I believed that it was raining.

In the first case, my reason for acting seems to be a feature of the world, whilst in the second it seems to be a feature of my psychology. Can these two different accounts of what my reason for acting was be reconciled within a single theory of what it is to be a reason? Most theorists think not; most theorists think that in one or the other of these cases we are just wrong about what my reason for acting was.

First, there are those theorists who take what happens when I am mistaken to be indicative of what happens when I'm not. They insist that, irrespective of whether or not it was raining, my reason for taking my umbrella was that I believed that it was raining (and not *that it was raining*). Theorists of this sort, so-called 'psychologists', suggest that our reasons for acting can only be features of our psychology.

Second, and in contrast, there are those theorists who take what happens when I'm *not* mistaken to be indicative of what happens when I am. They insist that, in both cases (i.e. regardless of whether or not it was raining), my reason for taking my umbrella was that it was raining (and not *that I believed that it was raining*). These theorists, whom I call 'deliberativists', insist that the consideration in light of which an agent acts is their reason for acting.

Third, there are those theorists, whom I call 'favourists', who do not seek to learn something from one case about the other; instead, they insist that our reasons for acting must count in favour of our actions. Like deliberativists, favourists argue that, when it was raining, my reason for taking my umbrella was that it was raining. However, unlike deliberativists, favourists insist that if I take my umbrella because of a mistaken belief that it is raining then I don't take it for a reason.

In contrast to all of these theorists, I think that the two different accounts of what my reason for taking my umbrella was can be reconciled within a single theory of reasons. I argue that what the fact that it is raining and the fact that I believe that it is raining have in common is that, in their respective cases, they each explain why it was rational for me to take my umbrella and why I took it. I suggest that it is in virtue of standing in those explanatory

relations to my action that those facts were, in their respective cases, my reason for taking my umbrella. More generally, I argue that there is a clear sense of what it is to be a reason there is for one to act, or a reason one has to act, or a reason for acting, according to which such reasons explain why it is, in a certain sense, rational for the agent to do the actions for which they are reasons. I call this account 'explanatory rationalism'.

# 1 Why do we need a new theory of reasons?

If there are already three, distinct, and popular theories of the relation between an agent's reason for acting and their action, why go looking for a fourth? Why do we need a new theory of reasons?

We need a new theory of reasons because the existing theories are, as Jonathan Dancy puts it, 'awkward in the sort of way that is usually due to bad theory. As Aristotle said, they leave one saying things that nobody would say unless defending a theory.' (2008a, 267) For instance, it is awkward to say, as the psychologist must, that my reason for taking my umbrella could never be that it was raining. And it is awkward to say, as the deliberativist must, that I could take my umbrella for the reason that it was raining, even though it wasn't. And it is likewise awkward to say, as the favourist must, that although I take my umbrella deliberately, purposefully and intentionally, I don't take it for a reason. And indeed, as I will show, these theories are awkward in yet other ways still.

I stress the awkwardness of these theories not because I take that to be the last word on their plausibility; clearly, one could just accept this awkwardness as a price that is worth paying for an otherwise convenient account of what it is to be a reason. My aim is rather to make clear that for each of these theories *there is a price that must be paid*, and that the price of each theory is sufficient to motivate the search for less costly alternatives. As Maria Alvarez puts it: *prima facie* paradoxical claims are not, 'a decisive argument against the views that generate them but [they] seem to favour considering whether there is a plausible alternative view that does not commit one to such claims.' (2016b, 11) So is there?

# 2 Pluralism about reasons

Well, here is an alternative that some theorists favour: perhaps there are just different senses of what it is to be a reason. For instance, perhaps there is a sense of what it is to be a reason in which reasons are what *psychologism* says they are, and perhaps there is a different sense of what it is to be a reason in which reasons are what *favourism* says they are. According to such an account, if I take my umbrella because I mistakenly believe that it is raining, there is a (psychologistic) sense in which I take it for a reason and a (favourist) sense in which I don't;

likewise (to pick a different expression), there is a sense in which I *had* a reason to take my umbrella and a sense in which I didn't. According to this *pluralist* theory of reasons, the awkwardness that *univocal* theories of reasons face is merely the result of conflating two different senses of a single expression; pluralist theories thus purport to be exactly the sort of plausible alternative that we are looking for. But are they?

Well, I agree that when I mistakenly believe that it is raining there is a sense in which I take my umbrella for a reason and a sense in which I don't; so, to that extent, I am persuaded that our eventual theory of reasons ought to be pluralist. However, I am not persuaded that we should rely on pluralism as a means of avoiding the problems that univocal theories of reasons face. In particular, it is not at all clear to me that whenever a univocal theory ends up entailing an awkward claim it is because it has conflated two different senses of a single expression. And, in the light of that scepticism, pluralism seems less like a way of avoiding awkwardness, and more like a capitulation to it.

# 3 Explanatory rationalism

Where does all this leave me? It leaves me searching for a univocal account of what it is to be a reason that doesn't force me to say awkward things (or, more modestly, that doesn't force me to say *the same* awkward things as these other theories). To that end, I present *explanatory rationalism*:

| **Reason expression** | **Explanatory rationalism** |
|---|---|
| For any *p, p* is a reason for *A* to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. |
| For any *p, p* is a reason for *A's φing*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and *p* makes *A's φing*, in some respect, worth doing. |
| For any *p, p* is a reason *A* has to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. |
| For any *p, p* is *A's* reason for *φing*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and explains (in the right way) why *A φ'd*. |

**Table 1:** Explanatory rationalism[1]

I will eventually argue that explanatory rationalism avoids all the problems that face other theories because it is able to reconcile the idea that agents always act for psychological reasons with the idea that they may sometimes also act for non-psychological reasons. But

[1] Two points are worth noting: First, I say that an action is *pro tanto* rational for *A* if and only if *A* takes it to be, in some respect, worth doing. See § (VI)1 for further discussion. Second, the categorisation schema used in Table 1 is unusual in so far as it allows for the possibility of distinguishing between each of the expressions listed (it is conventional to take at least some of these expressions to be co-extensive). I defend this approach, at length, in § (I).

before I reach that conclusion, I address what I take to be the major problem for explanatory rationalism.

## 3.1 The problem for explanatory rationalism

Explanatory rationalism claims, amongst other things, that an agent acts for the reason that *p* only if *p* explains *both* why they acted *and* why it was (*pro tanto*) rational for them to so act. This means that if explanatory rationalism is to be consistent with the claim that when I saw that it was raining my reason for taking my umbrella was that it was raining (as I intend it to be), then it must be possible for the fact that it was raining to explain *both* why I took my umbrella *and* why it was rational for me to take my umbrella.

There is, however, a well-rehearsed argument against this possibility. This argument says that facts about things that are external to an agent's mind (e.g. the fact that it is raining) can't explain why they did something or why it was rational for them to do it, since what an agent does, or what it is rational for them to do, only depends on their mind.

For instance, notice that even if it isn't raining, so long as I believe that it is raining I will still take my umbrella; and I will take it *because* I believe that it is raining. Further, notice that, given that I need to believe that it is raining in order to take my umbrella, even if I take my umbrella when it is raining I must still take it *because* I believe that it is raining. Thus, the argument goes, if the fact that I believe that it is raining can explain why I took my umbrella *whether or not it is raining*, then the fact that it is raining does no real work in explaining why I took my umbrella, and, therefore, it does not explain why I took it.[2]

This argument is a problem for any theory that says that an agent's reason for acting could be something other than a feature of their psychology (as explanatory rationalism, favourism and deliberativism all do), and is thus probably the motivating argument for psychologism about agents' reasons for acting. Thus, in order to save explanatory rationalism, I need to say how it is that (in spite of the argument above) the fact that it is raining can explain both why I took my umbrella, and why it was rational for me to take it.

## 3.2 My response to the problem

My response to this problem is to say that the fact that it is raining explains why I took my umbrella by explaining why I believed that it was raining, which in turn explains why I took my umbrella. Likewise, I suggest, the fact that it was raining explains why it was rational for me to take my umbrella by explaining why I believed that it was raining, which in turn explains why it

[2] This argument can likewise be applied to the explanation of why it was rational for me to take my umbrella.

was rational for me to take my umbrella. I argue that the mistake in the argument above is that it rejects the possibility of such *distal* explanations.

This response relies on the transitivity of the explanatory relations involved: if the explanatory relations between the fact that it was raining, the fact that I believed that it was raining, and the fact that I took my umbrella weren't transitive, then the fact that it was raining would not explain why I took it.

However, explanatory relations are not always transitive. For example, if I mistake the spray from a sprinkler for rain, then the fact that the sprinkler is spraying in front of my window explains why I believe that it is raining, which explains why it is rational for me to take my umbrella. But, the fact that a sprinkler is spraying in front of my window does not explain why it is rational for me to take my umbrella, in spite of the explanatory chain connecting the two facts. So, the explanatory relations involved aren't transitive with one another.

Why is it that the transitivity of explanation fails in this case, but apparently succeeds when I actually see rain? That is, given that the fact that the sprinkler was spraying in front of my window and the fact that it was raining both (in their respective cases) explain why I believed that it was raining, why is it that (as I have suggested) only the latter explains why it was rational for me to take my umbrella? It is because, I will argue, the explanatory relationships in the case of the latter, unlike the former, are all of a certain, transitive sort. In particular, I argue that there is a *mysterious*, non-causal explanatory relation that obtains, *inter alia*, between the fact that *p* and the fact that an agent believes that *p* when an agent knows that *p*.

I thus argue that, in the case when I saw rain, the fact that it was raining explains why it was rational for me to take my umbrella because it is *mysteriously* related to the fact that I believed that it was raining (which in turn explains why it was rational for me to take my umbrella). And I argue that, conversely, in the case in which I was mistaken, the fact that the sprinkler was spraying in front of my window does not explain why it was rational for me to take my umbrella because it is not mysteriously related to the fact that I believed that it was raining (it is merely causally related to it). Moreover, I argue, in both of these cases the fact that it appeared to me as though it was raining is mysteriously related to the fact that I believed that it was raining, so that in both of these cases the fact that it appeared to me as though it was raining explains why it was rational for me to take my umbrella (and, indeed, explains why I took it).

# 4 My theory of reasons

Thus, according to explanatory rationalism, when I saw that it was raining, my reasons for taking my umbrella include the fact that it was raining, the fact that it appeared to me as though it was raining, and the fact that I believed that it was raining. In contrast, when I merely saw the spray of the sprinkler, my reasons for taking my umbrella do not include the 'fact' that it was raining (not least because it wasn't raining) nor do they include the fact that the sprinkler was spraying in front of my window, but they nonetheless include the fact that it appeared to me as though it was raining, and the fact that I believed that it was raining.

Unlike any other univocal theory of reasons, explanatory rationalism is thus consistent both with the claim that agents always act for psychological reasons and with the claim that they sometimes act for non-psychological reasons. And, indeed, it is precisely this that means that explanatory rationalism avoids the particular awkward claims that other theories face, and which, more generally, makes it immune to many of the challenges one would normally level against such theories. Thus, I argue, explanatory rationalism is the best univocal theory of reasons.

However, since, as I noted, I am persuaded that there may be two distinct senses to each reason expression, my own theory of reasons is what I call 'new pluralism'. New pluralism says that one sense of every reason expression is explanatory rationalist, while the other sense is favourist; and this allows me to say that, for instance, when I take my umbrella because I mistakenly believe that it is raining, there is a sense in which my reason for taking my umbrella is that I believe that it is raining, and there is a sense in which I take my umbrella for no reason. The virtue of new pluralism over existing pluralist theories is that it does not rely on the plurality of senses to avoid the awkwardness that univocal theories face – explanatory rationalism already avoids that awkwardness on its own.

# 5 That which I pass over in silence

There are a few topics about which this discussion makes no claim. The first of these is the ontology of reasons. For the most part I will talk as though reasons are propositions, if not facts, however, this is mostly for convenience. Whilst there has been some debate between those who take reasons to be facts (or states of affairs) and those who take them to be psychological states, I make no particular claim about that. My theory is not about *what reasons are*; it is about the relation in which reasons stand to the actions for which they are reasons. Since I take reasons to stand in explanatory relations to the actions for which they are

reasons, if you accept my account then it will be (at least in part) your view on the ontology of *explanantia* that determines what the ontology of reasons is. And about that I make no claim.

This leads me to the second topic about which I make no claim: what the correct analysis of explanation is. Whilst I discuss explanation in some depth, I make no claims about the specific relations that underpin explanatory relations, and what the conditions for them are. However, it may be that some of the purported instances of explanation that I cite are incompatible with some accounts of explanation. To my knowledge, there are no instances of this kind that would be of serious concern for my argument, and it is my belief that an alternative construal of my argument could be made whatever one's theory of explanation. However, I may be wrong; there may be accounts of explanation that are inconsistent with what I want from my theory of reasons, in which case, so be it.

Thirdly, I will also leave the question of whether or not desires are reasons well alone. Of course, we regularly say things like: 'my reason for going to the gym is that I want to get fit'; and, 'my reason for going to the shops is to buy eggs'. Whilst I think that explanatory rationalism has something to say about how we interpret these sentences, I won't discuss it here, and I won't mention such sentences further.

Finally, I will not discuss reasons for belief. Many hold that reasons for belief should be analysable in the same sort of way as practical reasons, and this is a view that I share. It would be tempting, then, to extend explanatory rationalism into an analysis of what it is to be a reason to believe something, and, indeed, I think that such an account could be illuminating. However, I don't discuss that here.

# 6 An overview of this discussion

The structure of this discussion is as follows. In § (I), I set out my approach to talking about theories of reasons. In particular I categorise theories of reasons according to the claims they make about each of the following expressions: 'the reasons there are for one to act'; 'the reasons for or against acting'; 'the reasons one has to act'; and 'one's reason for acting'. This is at odds with the prevailing tendency to sort reasons into 'normative' and 'motivating'; however, for reasons that I will make clear, I prefer to eschew that terminology in my categorisation schema.

The discussion of § (I) highlights three claims about reasons that are probably the most widely held (though not necessarily by the same theorists). These are:

> *Favourism about reasons to act*: For any *p, p* is a reason for *A* to *φ* if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing.
>
> *Psychologism about the reasons for which we act*: For any *p, p* is a reason for which *A φ*'d if and only if *p* is a feature of *A's* psychology that rationalises *φ*ing and explains (in the right way) why *A φ*'d.
>
> *Deliberativism about the reasons for which we act*: For any *p, p* is a reason for which *A φ*'d if and only if *p* is a consideration in light of which *A φ*s.

In §§ (II), (III) and (IV), respectively, I show that each of these accounts is inconsistent with a number of *prima facie* reasonable claims. The problems set out in these chapters are mostly alternative construals of arguments that have already been made against each of these accounts, and they form the basis of my claim that all existing theories of reasons are, in some sense, *awkward*.

In § (V), I consider whether we should just adopt pluralism as a solution to the awkwardness of univocal theories. I conclude that we should not: pluralism, at least of the conventional sort, falls short of what we need from a new theory of reasons.

In § (VI), I set the agenda for the remainder of the discussion. I introduce *explanatory rationalism*, and I give an overview of the main problem for it (highlighted in the discussion above), which I call 'The Explanatory Exclusion Problem'. I provide an outline of my approach to discussing, and, ultimately, solving this problem, which is as follows.

In § (VII), I make some assumptions about the structural principles and logical properties of explanation. While my discussion relies on characterising explanation in this way, I do not think that either my solution to The Explanatory Exclusion Problem, or my theory of reasons more generally, depends on explanatory relations being so characterisable.

In § (VIII), I use the assumptions of the preceding chapter to provide a formal construal of The Explanatory Exclusion Problem. In particular, I show how its conclusion (that facts about the world external to an agent's mind cannot explain the agent's actions) can be arrived at from two seemingly trivial claims about what explains an agent's action when they act in error or in ignorance, together with five seemingly plausible principles of explanation.

In § (IX), I show how The Explanatory Exclusion Problem can also be used to argue that facts about the world external to an agent's mind cannot explain why it is rational for them to act.

In § (X), I consider two popular accounts of how, for instance, the fact that it is raining explains why I took my umbrella. The first of these accounts says that it does so *elliptically*; the second says that it does so *directly*. I argue that both of these accounts are flawed.

In §§ (XI) & (XII), I develop my account of how the fact that it is raining explains why I took my umbrella; I say that it explains it *indirectly*. My argument proceeds in two stages. First, in § (XI), I argue that we should reject the conclusion of The Explanatory Exclusion Problem because it is based on a false principle of explanation, *the exclusion principle*. The exclusion principle implies that only the most *proximal* explanations of some *explananda* explain it; but this is mistaken – I argue that most of the explanations we are interested in are, to some extent, *distal* explanations. Then, in § (XII), I show how that insight helps inform the account of how the fact that it is raining can explain why I took my umbrella. Specifically, I suggest that the fact that it is raining is a *distal* explanation of my action; it explains why I took my umbrella by explaining a more *proximal* explanation of why I took my umbrella (namely, the fact that I believed that it was raining).

In § (XIII), I suggest that the same account explains how it is that, for instance, the fact that it is raining can also explain why it is rational for me to take my umbrella. That is, the fact that it is raining explains why I believe that it is raining, which in turn explains why it is rational for me to take my umbrella.

However, I note, this does not mean that if an agent's belief explains why it is rational for them to do some action, then anything that explains why they have that belief will also explain why it is rational for them to do that action. That is, I note, not all explanatory relations are transitive. I then set the stage for the remainder of the discussion, which seeks to provide an account of when the explanatory relations involved are, and when they aren't, transitive.

My account proceeds in two stages. First, in § (XIV), I introduce the *mystery* relation. I argue that the mystery relation is a non-causal, transitive, explanatory relation that relates: the belief that *p* to some justification for it when that belief is justified; the belief that *p* to the fact that *p* when the belief that *p* is knowledgeable; a justification for the belief that *p* to the fact that *p* when that justification affords the opportunity for knowledge; and an action to some belief that explains why that action is rational when that action is done intentionally. Then, in § (XV), I argue that the mystery relation is transitive with the non-causal explanatory relation involved in explaining why some action is rational, whereas merely causal explanatory relations are not. This is why facts that merely causally explain our beliefs (such as the fact that the sprinkler is

spraying in front of my window, or facts that deviantly cause our beliefs) do not explain why it is rational for us to act.

Finally, in § (XVI), I revisit explanatory rationalism. I show how it responds to the problems faced by other theories. I suggest that it is to some extent immune from the conventional procedures for generating problems for theories of reasons because it is consistent both with the claim that agents always act for psychological reasons and with the claim that they sometimes act for non-psychological reasons. In light of these arguments, I suggest that explanatory rationalism is the best univocal account of what it is to be a reason.

I conclude by returning to the intuition that, when someone acts on a false belief, there is a sense in which they act for a reason and a sense in which they don't. Since I share this intuition, I advocate *new pluralism*, which says that explanatory rationalism tells us one sense of what it is to be a reason, whilst favourism tells us the other. I show that new pluralism does not face the same weaknesses as other forms of pluralism because whenever an agent acts for a reason in the favourist sense, they also act for a reason in the explanatory rationalist sense.

# Chapter Summary

**(I) ON THEORIES OF REASONS**

In which I say how we should categorise theories of reasons. I argue that, if our categorisation schema is to capture at least the main theories of reasons, then it should by allow for the possibility of as many kinds of reason as there are reason expressions. I say that instead of distinguishing between psychological and anti-psychological theories we should categorise theories of reasons according to what they say, for each reason expression, about the relation between the reasons picked out by that expression and the actions for which they are reasons. And I say that we should eschew the terminology of 'normative' and 'motivating' reasons in our categorisation schema and in our theorising, because, as they are standardly defined, they make substantive claims about what it is to be a reason that it is properly the business of a theory of reasons to determine, and that, moreover, those claims systematically disadvantage 'psychological' theories of reasons. To frame the discussion of practical reasons in terms 'normative' and 'motivating' reasons is, I suggest, a Trojan horse that earnest theorising ought to reject.

**(II) REASONS TO ACT THAT MAKE ACTIONS WORTH DOING**

In which I show what it costs to think that if there is a reason to do some action then that action is, in some respect, worth doing. I show how 'favourism about reasons to act' (which entails that reasons to act make actions worth doing) clashes with some prima facie reasonable claims about a case in which someone acts on a false belief. I set out which claims the proponent of this view must choose between rejecting and I categorise the common choices from the literature.

**(III) ACTING FOR PSYCHOLOGICAL REASONS**

In which I show what it costs to think that the reason for which an agent acts is always a feature of their psychology. I show how 'psychologism about the reasons for which we act' clashes with some prima facie reasonable claims. In particular, I show that is inconsistent with the idea that we are often able to act for reasons that make our actions morally worthy and, more generally, worth doing.

**(IV) ACTING FOR WHAT YOU BELIEVE**

In which I show what it costs to think that a reason for which an agent acts is the content of the belief they acted on. I show how 'deliberativism about the reasons for which we act' clashes with some prima facie reasonable claims about the factivity of reasons, the explanatory power of the reasons for which we act, the factivity of explanation and what an agent's reasons for acting are in Gettier cases. I set out which claims the proponent of this view must choose between rejecting.

**(V) ON THE PLURALITY OF REASONS**

In which I explain what a pluralist theory of reasons is and why 'going plural' is not a panacea. I suggest that a given reason expression could have more than one sense, and I show how we can accommodate theories of reasons that accept that idea, i.e. pluralist theories of reasons, in our categorisation schema. I discuss some examples of pluralist theories from the literature. I show how pluralist theories can solve some of the problems discussed in the previous chapters. I explain why pluralism is not, however, enough, and I suggest that our investigation should go beyond favourism, psychologism and deliberativism.

**(VI) A NEW FAMILY OF CLAIMS ABOUT REASONS**

In which I set out a new family of claims about reasons, and introduce the major challenge to it. I define 'pro tanto rational' actions as actions that an agent takes to be, in some respect, worth doing. I set out a new family of claims about reasons, explanatory rationalism, which says that all practical reasons explain why the actions for which they are reasons are pro tanto rational. I introduce the major challenge for explanatory rationalism, The Explanatory Exclusion Problem, which argues that only features of an agent's psychology could explain either why they do something or why it was rational for them to do it. I set out the program for the forthcoming chapters.

**(VII) WE NEED TO TALK ABOUT EXPLANATION**

In which I make some assumptions about explanation. I say what I mean by 'explains' and I state that I will talk as though explananda are facts and explanantia are propositions (whether or not they are). I distinguish two sorts of explanatory relation, 'fully explains' and 'partially explains', where a full explanation is sufficient for the truth of the fact that it explains and a partial explanation is an element (or subset) of a full explanation, and I make some assumptions about the logical properties of these relations. Lastly, I say that some fact is 'overexplained' just in case there are two genuinely different full explanations of that fact.

**(VIII) THE EXPLANATORY EXCLUSION PROBLEM**

In which I set out The Explanatory Exclusion Problem, which is, in some form or another, the motivating argument for psychologistic theories of reasons. I provide a formal construal of the Problem, showing how it results from two seemingly trivial claims about what explains an agent's action when they act from error and from ignorance together with five seemingly plausible principles of explanation. I show how the Problem implies that I did not congratulate my friend because she had won an award, but only because I thought she had.

**(IX) OTHER USES FOR THE EXPLANATORY EXCLUSION PROBLEM**

In which I show how The Explanatory Exclusion Problem can be used to arrive at some other conclusions that are inconvenient for explanatory rationalism. I set out the general form of the Problem, followed by the general form of the argument for the first premise of the Problem. I show the Problem can be used to argue that the fact that I read that my friend had won an award does not explain why I congratulated her, and that neither that fact, nor the fact that she had won an award, can explain why it was pro tanto rational for me to congratulate her.

**(X) HOW NORMATIVE REASONS DON'T EXPLAIN**

In which I reject two accounts of how normative reasons explain. I re-introduce talk of normative reasons, defining them as things that make actions, in some respect, worth doing. I ask how it is that we manage to explain our actions when we say that we acted because of a normative reason there was to act; for instance: how is it that I explain why I took my umbrella when I say that I took it because it was raining? I suggest that the fact that it was raining explains why I took my umbrella either 'elliptically', 'directly' or 'indirectly'. I note that which answer one accepts will depend on one's response to The Explanatory Exclusion Problem: elliptical theorists accept the conclusion of the Problem, direct theorists reject the first premise, and indirect theorists reject the second. I set out the problems with elliptical and direct theories.

**(XI) THE EXCLUSION PRINCIPLE IS FALSE**

In which I show that the exclusion principle is false. I provide two counterexamples to the exclusion principle, one involving causal explanation and another involving non-causal explanation. I suggest that they are counterexamples because in each case the purportedly excluded fact explains the explanandum by explaining something that, in turn, explains the

explanandum. I suggest that the problem with the exclusion principle is that it discriminates against all but the most proximal explanations of any given explanandum, and that this is problematic at least partly because we are typically interested in more distal explanations. I explain where our reasoning went wrong and which full explanation an apparently excluded fact is part of.

**(XII) EXPLAINING WHY WE ACT**

In which I say how normative reasons (and the appearance of them) explain why we act. I suggest that normative reasons explain an agent's action by explaining their belief that, in turn, explains the agent's action. I suggest that they explain an agent's belief by explaining the appearance of them that, in turn, explains the agent's belief. I set out the implications of this view for explanatory rationalism and for anti-psychological theories of reasons more generally.

**(XIII) EXPLAINING WHY IT IS RATIONAL TO ACT**

In which I say when something explains why it's rational to act, and when it doesn't. I suggest that normative reasons or appearances explain why it is rational to act only if they explain those beliefs that in turn explain why it is rational to act. I note that it is tempting to infer that if an agent's belief explains why it is rational for them to do some action then whatever explains that belief also explains why it is rational for them to do that action. I show how that inference leads to an apparent dilemma for explanatory rationalism. I counsel against that inference, by noting that different kinds of explanatory relations may not be transitive with each other. I then set out the task ahead: showing that the explanatory relations concerned are transitive when, and only when, explanatory rationalism needs them to be.

**(XIV) THE MYSTERY RELATION**

In which I introduce the mystery relation. I suggest that a mysterious, non-causal relation obtains between a belief and the justification that it is based on when that belief is justified. I argue that the mystery relation must be non-causal, because, as deviant causal chains demonstrate, a merely causal relation between a belief and some justification for it is not sufficient for that belief to be justified. I suggest that this exact same mysterious relation relates: the belief that p to the fact that p when the belief that p is knowledgeable; a justification for the belief that p to the fact that p when that justification affords the opportunity for knowledge; and an action to some belief that explains why it is rational when that action is done intentionally. I argue, furthermore, that this mystery relation is a transitive, explanatory relation.

**(XV) MYSTERY RELATIONS AND WHY IT IS RATIONAL TO ACT**

In which I say that mystery relations are transitive with the explanatory relation involved in explaining why it is rational. I label the sort of explanatory relation that obtains between (i) the fact that I believe that it is raining and (ii) the fact that it is pro tanto rational for me to take an umbrella, the 'E'-relation'. I argue that the mystery relation is transitive with the E'-relation. I show how this accords with our intuitions in some of the examples already considered.

**(XVI) A NEW THEORY OF REASONS**

In which I set out my theory of reasons. I discuss what explanatory rationalism says about the application of each reason expression to the case where I take my umbrella having seen that it is raining. I show how explanatory rationalism solves the problems faced by other theories. I suggest that the best theory of reasons is a pluralist theory of reason that combines explanatory rationalism and favourism; I call this theory 'new pluralism'. I show how explanatory rationalism enables new pluralism to meet the main challenge to pluralist theories.

# (I)

# On Theories of Reasons

*In which I say how we should categorise theories of reasons. I argue that, if our categorisation schema is to capture at least the main theories of reasons, then it should by allow for the possibility of as many kinds of reason as there are reason expressions. I say that instead of distinguishing between psychological and anti-psychological theories we should categorise theories of reasons according to what they say, for each reason expression, about the relation between the reasons picked out by that expression and the actions for which they are reasons. And I say that we should eschew the terminology of 'normative' and 'motivating' reasons in our categorisation schema and in our theorising, because, as they are standardly defined, they make substantive claims about what it is to be a reason that it is properly the business of a theory of reasons to determine, and that, moreover, those claims systematically disadvantage 'psychological' theories of reasons. To frame the discussion of practical reasons in terms 'normative' and 'motivating' reasons is, I suggest, a Trojan horse that earnest theorising ought to reject.*

Within the domain of practical reasons, it is common to distinguish two kinds of reason: normative and motivating. It is typical[1] to then categorise theories of reasons according to whether or not they take reasons of each kind to be (exclusively) features of the agent's psychology. If we say that a theory is 'psychological' with respect to a given kind of reason if it says that a reason of that kind is always a feature of the agent's psychology, and 'anti-psychological' if it denies that view, we have the following, 'standard' categorisation of theories of reasons:

| **Theory** | **Normative reasons** | **Motivating reasons** |
|---|---|---|
| 'The Received View'[2] | Anti-psychological | Psychological |
| 'Anti-psychologism' | Anti-psychological | Anti-psychological |
| 'Extreme psychologism'[3] | Psychological | Psychological |

**Table I-1:** The 'standard' categorisation of theories of reasons

Despite its ubiquity, this categorisation schema is, as I will argue, only useful within the confines of a fairly narrow theoretical framework. In particular, it is not well suited to the task (for which it may well not have been intended) of distinguishing many of the main theories of

[1] (E.g. Dancy 2000; Sandis 2013; O'Brien 2015)

[2] This view is the target of Dancy's criticism in his *Practical Reality* and is what Mitova (2015) calls the 'Standard Story'. It is typically attributed to Smith (1987), however, in § (V) I will challenge this attribution.

[3] 'Extreme psychologism' is a name that I have borrowed from Mitova (2015), who defends that view.

reasons from one another, because: (i) it only allows for two kinds of practical reason, while some prominent theories allow for more; (ii) the categorisation of theories into psychological and anti-psychological cannot discriminate between distinct anti-psychological theories; and (iii) given the way that 'normative' and 'motivating' reasons are now standardly defined, assuming that they are picked out by *any* given reason expression involves a theoretical commitment that at least some theories of reasons would and should reject.

In criticising this categorisation schema I mean to criticise no work in particular. It may be that the schema, so represented, is a straw man in so far as no one holds it in its entirety (or at least not for the purpose of sorting different theories of reasons). Nonetheless, I take it that everything that I critique is held by at least some, and some of that which I critique is held by many – so, since my discussion is of the constituent parts of the schema, whether or not the complex is generally endorsed is neither here nor there. If it turns out that everyone agrees that the standard schema is of no use in categorising distinct theories of reasons then what the discussion that follows will have done is demonstrate why that is the case, as well as offering an alternative that is better suited to that end.

To re-iterate, my aim, in what follows, is to demonstrate that what I have called the 'standard' schema is not well suited to the task of discriminating between different theories of reasons, and to propose an alternate categorisation schema that is better suited to that task.

# 1 How many kinds of reason are there?

In general, if we use one expression to refer to one thing and a different expression to refer to something else, we have a *prima facie* basis for thinking that the two things are of different kinds. Of course, it may turn out that two different expressions are used to pick out the same kind of thing (e.g. 'a bachelor' and 'a single man'), but *before* we determine what kind of thing they pick out, a difference in the form of any two expressions gives us a *prima facie* basis for thinking that they pick out different kinds of thing.

Conversely, if we use the same expression to refer to two different things, we have a *prima facie* basis for thinking that the different things are of a common kind. Of course, it may turn out that the same expression can be used to pick out *different* kinds of thing (as in the case of homonyms like 'a bat' (an animal) and 'a bat' (an item of sports paraphernalia)), but before we determine that that is the case, the sameness of the expression gives us a *prima facie* basis for thinking that the things that it picks out are all of a common kind.

So, when we want to work out how many kinds of reason there are, a good way to start is to look at the different expressions that we use to pick out things that we call 'reasons'. Here are what I will take to be the main ones:

- Reasons there are to act;
- Reasons for and reasons against acting;
- Reasons one has to act;
- One's reason for acting; and
- Reasons why.

All of these expressions pick out *reasons* of one kind or another. Each expression has a different form, so, I suggest, at least *before* we determine what kind of reason they each pick out, we have a *prima facie* basis for thinking that they each pick out *different* kinds of reason. Moreover, because homonyms are the exception, not the rule, we have a *prima facie* basis for believing that they each pick out only one kind of reason. Of course, further analysis may yield the finding that some of these expressions pick out the same kind of reason, or that some pick out more than one kind of reason, but that should be a conclusion of our theorising, not an assumption with which we begin. So, since we have five different *reason expressions*, we have a *prima facie* basis for thinking that there are five different kinds of reason – one corresponding to each reason expression.

Some clarifications, for the avoidance of doubt: firstly, I take two kinds of reason, kind *A* and kind *B*, to be different kinds of reason if and only if the conditions for being a reason of kind *A* differ from the conditions for being a reason of kind *B*. Secondly, saying that there are different kinds of reason does not mean that one and the same thing cannot be a reason of each kind, but nor does it mean that they can. Rather, it is for your theory of reasons to determine whether or not reasons of one kind can also be reasons of another kind; and it will do so by telling you what the conditions for being a reason of each of those kinds are. Lastly, two reason expressions pick out reasons of the same kind if and only if they are co-extensive.

## 1.1 Reason expressions in the standard schema

The standard schema assumes that there are just two kinds of reason – this is the assumption I wish to criticise in this section. The standard schema also makes assumptions about the conditions under which something is a reason of either kind, which further restricts its usefulness, but I will return to that issue in § 4.

Given that each reason expression picks out reasons of one kind or another, in assuming that there are only two kinds of reason, the standard schema must assume that several of the reason expressions considered pick out reasons of the same kind. And that is indeed so: the standard schema typically assumes (either implicitly or explicitly) that the first three reason expressions ('reasons there are to act'; 'reasons for/against acting'; and 'reasons one has to act') all pick out reasons of one kind, while the latter two expressions ('one's reason for acting' and 'reasons why one acts') pick out reasons of the other kind.

The standard schema may well be correct about the co-extensivity of some of these expressions – but that is a conclusion to be argued for, it is not something to be assumed before one has begun theorising in earnest. One reason not to make this assumption from the outset is that we have a *prima facie* basis for thinking that the different expressions aren't co-extensive – namely, the fact that they are different expressions.

However, the differences between the expressions notwithstanding, assumptions of their co-extensivity might be tolerable if the equivalence between particular expressions were obvious, or at least un-contentious – but that is certainly not the case. Indeed, as we shall see, several theories of reasons already distinguish between expressions that the standard schema takes to be co-extensive, and those theories consequently escape categorisation within that schema.

I want to make the case for a categorisation schema that allows for the possibility of distinguishing between each of the reason expressions I list. I also want to make clear where a theory of reasons is *required* to draw a distinction between different expressions. And, lastly, I want to say which of these reason expressions we should and which we should not include in our categorisation schema. The following sections discuss each expression, with a view to achieving each of these aims.

## 1.2 Reasons there are to act

There is a reason for me to go for a swim. To avoid controversy I won't say what it is – but I take it to be uncontroversial to say that if (and only if) there is a reason for me to swim then something[4] is the reason that there is. I take it to be likewise uncontroversial to say that if something is a reason *there is* for me to swim then it is a reason for me to swim.[5]

[4] I make no assumption (yet) about the ontology of that something – it could be a fact, a state of affairs, a mental state or whatever you like.

[5] These remarks are obviously intended to generalise beyond me and my swimming. But, for the avoidance of doubt, I take the statement that 'something is reason there is for *A* to *φ*' to be logically equivalent to 'that same thing is a reason for *A* to *φ*' and 'there is a reason for *A* to *φ*'. I take the difference in the forms of words here to be so slight as to preclude effective distinction – so our

Here is a seemingly 'pre-theoretic' observation we can make about the kind of reason picked out by the expression 'the reasons there are to act': something can be a reason for *A* to *φ* even if *A* does not *φ*. For instance, there can be a reason for me to swim even if I don't swim. That is, it does not impinge on the reason-hood of a reason to act if one does not do the action that it was a reason to do.

I suggest we say that if the reason-hood of reasons of a certain kind is not dependent on the occurrence of the action for which they are a reason, then reasons of that kind are 'independent' of the actions for which they are reasons.[6] Thus, reasons to act are *independent* of the actions for which they are reasons.

## 1.3 Reasons for and reasons against acting

I'm debating whether or not to swim: there are reasons for and reasons against my swimming[7] – don't worry about what they are. Like reasons there are to act, reasons for or against acting are *independent* of the actions for which they are reasons: something can be a reason for swimming (or a reason against swimming) even if I don't (or do) swim.

While they differ in their form, I have grouped 'reasons for' and 'reasons against' because there seems to be a clear sense in which their meanings are inter-related – they are opposites. So whatever we learn about the kind of reason picked out by one will tell us something about the kind picked out by the other. As is typical in the literature, I will mostly restrict my discussion to the 'reasons for' expression.

As we noted in § 1.1, it is typically assumed that 'reasons there are *to* act' and 'reasons *for* acting' are co-extensive expressions (it is also typical to assume that 'reasons *not to* act' and 'reasons *against* acting' are likewise co-extensive). I think this is a mistake – I think that the prepositions matter to the meaning, and that the two expressions pick out different kinds of reason. However, I am not arguing for this claim at this stage; for now, I am only arguing that our categorisation schema *should allow for the possibility of difference between them*.

That is, at this stage I neither accept nor deny that 'reasons there are to act' and 'reasons for acting' always pick out the same kind of reason: all that I am saying is that, in spite of the fact that they are both *independent* of the actions for which they are reasons, the difference in the

---

categorisation schema need not distinguish between them. Nonetheless, I state these assumed equivalences with a view to making my assumptions clear.

[6] This will contrast them with those kinds of reason (introduced in § 1.5) whose reason-hood *is* dependent on the occurrence of the action for which they are reasons, which I call 'dependent' reasons.

[7] Again, if there are reasons for swimming then something is a reason for swimming, and if there are reasons against swimming then something is a reason against swimming. And again, these equivalences generalise beyond me swimming.

form of these reason expressions gives a *prima facie* basis for thinking that they pick out different kinds of reason. Moreover, since I eventually want to advocate a theory that *does* distinguish between the reasons picked out by these expressions, if my theory is to be represented within my own categorisation schema, I had better allow for the possibility of such a distinction.

## 1.4 Reasons one has to act

I have a reason to swim. Again, I take it to be uncontroversial to say that if I have a reason to swim then something is the reason that I have to swim.[8] And, like reasons there are for me to swim (and reasons for swimming), reasons that I have to swim are *independent* of the actions for which they are reasons: one can have a reason to swim (or eat, or go to the shops) without doing so.

As we have seen, it is typical to take the expressions 'the reasons one has to act' and 'the reasons there are to act' to be co-extensive. As it happens, I share this view, but we nonetheless should not start with this assumption, much less incorporate it into our categorisation schema, for the following reasons: firstly, as I've already stressed, the difference between the forms of the expressions gives us a *prima facie* basis for thinking that they pick out different kinds of reason, so it is odd, if not counter-productive, to start (as the standard categorisation schema does) by assuming that they don't. Secondly, and more importantly, several authors (e.g. Hornsby 2008; Schroeder 2008; Comesaña and McGrath 2014) reject the view that these expressions are co-extensive; and their theories consequently escape categorisation within the standard schema. And thirdly, as we shall see, assuming that there is no distinction between the kinds of reasons picked out by these expressions conceals a possible solution to one of the main problems for the dominant account of the reasons there are to act.[9]

So, in spite of the fact that the kinds of reason picked out by the expressions, 'reasons one has to act' and 'reasons there are to act', are both *independent* of the reasons for which they are actions, we nonetheless have good grounds for using a categorisation schema that allows at least for the possibility of distinguishing between the kinds of reason picked out by these expressions.

[8] This makes no assumption about the ontology of that thing (in particular – saying that something *is* the reason I have to swim does not assume that that thing need be a fact rather than, say, an intentional object).

[9] See the discussion of the 'The Rational Action Problem' in § (II)4.1.

## 1.5 One's reason for acting

I think about how swimming will improve my mood, how I will sleep better if I swim, and other things besides and, after some deliberation, I decide to swim. I head off and do it: I swim for a reason. As ever, I won't say what my reason for swimming was.[10]

Some, hopefully, un-contentious equivalences: I swam for a reason if and only if something was my reason for swimming, and something was my reason for swimming if and only if (i) I swam for that reason and (ii) it was the reason *for which* I swam.[11] The latter remark is perhaps the most contentious (but hopefully not especially so) of the equivalences I have drawn, in so far as I take 'my reason for swimming' to be equivalent to 'the reason for which I ran' in spite of the obvious difference in the form of these expressions. It is certainly in principle possible to distinguish between them, however, I have elected not to on the basis that I know of no theory that does (so the equivalence is seemingly uncontentious) and so, for the sake of brevity, I assume no distinction from the outset. Perhaps I am not following my own instruction – but my assumption is at least plain, so that those who disagree with it may revise my schema knowing as much.

While the standard categorisation schema draws no distinction between the reason expressions discussed in the previous three sections, it does, as I have noted, distinguish between those reason expressions and 'the agent's reason for acting'. Now, while I want us to abandon the standard categorisation schema, I agree with its implicit verdict that there are unequivocal grounds for differentiating the kinds of reason picked out by 'the reason for which the agent acts' from the kinds picked out by the expressions already considered, which go beyond a mere difference in form.[12] What do I take those grounds to be?

Consider this: something cannot be one's reason for *φ*ing unless one *φ*s. That is, I can't swim for a reason unless I swim, so something can't be the reason for which I swam unless I did indeed swim. And this is true even if the action is in the future: something might *now* be the reason for which I *will* swim, but it is only so if I do indeed swim (if I don't go on to swim then it isn't *now* the reason for which I *will* swim). So, unlike the reason expressions already considered, the reason-hood of the reasons picked out by this expression are not independent

[10] In particular, the story of my deliberative process need lead us to no conclusions about what my reason for swimming was.

[11] Again, these remarks should generalise beyond me and my swimming, but, for the avoidance of doubt, I take the statement that 'something was *A*'s reason for *φ*ing' to be logically equivalent to 'that same thing is a reason for which *A φ'*d' and '*A φ'*d for a reason'.

[12] I call this its 'implicit verdict' only because the standard approach does not generally take a difference in form to be sufficient grounds for allowing the possibility of a distinction – as the failure to distinguish between the reason expressions already considered demonstrates.

of the actions for which they are reasons – they are, let us say, 'dependent' on them, in the sense that the reason-hood of such reasons depends on the occurrence of the action for which it is a reason.

So, the kind of reason picked out by the expression 'the reasons for which *A φ*s' is distinct from the kind(s)[13] picked out by the reason expressions of the previous sections (because the conditions for being a reason of the former kind differ from the conditions for being a reason of the latter kind(s)).[14] This means that not only must our categorisation schema allow for the possibility of distinguishing this sort of reason from those already considered, but any theory of reasons should also make clear that this expression *does* pick out a reason of a different kind.

## 1.6 Reasons why

There was a reason why I swam. Like my reasons for swimming (or the reasons for which I swam), the reasons why I swam are *dependent* on that which they are a reason for: if there is a reason why I swam then I must have swum.

As we have noted, the standard schema does not systematically distinguish between the kinds of reason picked out by 'the reason why someone acts' and 'the reason for which they act.' Nonetheless, the difference in their forms still gives us a *prima facie* basis for distinguishing the two expressions, and this is despite the fact that they both pick out a kind of reason that is *dependent* on that which it is a reason for. Moreover, recent scholarship has forcefully made a case for distinguishing between the kinds of reason picked out by the two expressions that goes beyond the mere difference in their form.[15]

Firstly, we can readily observe that, in spite of their both being *dependent* reasons, something can be a reason why an agent acts without being their reason for acting. For example, the fact that I'm chopping onions may be a reason why I'm crying without being my reason for crying (I am not crying for a reason, though there is a reason why I am crying). Likewise, the fact that Anthony was given a posthypnotic suggestion might be a reason why he drinks vinegar, without being his reason for drinking it, and that is so even if he drinks it for a reason. One

[13] The parenthetical pluralisation is to note that these reason expressions may or may not pick out different kinds of reason.

[14] We should note that this is not to say that one and the same thing cannot be a reason of each kind (as I already remarked in the clarifications at the end of § 1.1); it is just to note that the expressions pick out reasons of different kinds. That is, to the extent that a single reason can be picked out by the expression 'the reason for which they acted' and, for instance, 'a reason they had to act' (and, at this stage, I assume nothing about whether or not they can) it is because that which is picked out is both a *dependent* and an *independent* reason – it is two different kinds of reason.

[15] See, in particular, Alvarez (2010) but also Audi (2001), Dancy (2000) and Hieronymi (2011).

need not, I think, accept any particular theory of reasons in order to accept the truth of these claims and, moreover, their truth already provides a basis for distinguishing between the kinds of reason picked out by the expressions 'the reasons why one acted' and 'the reasons for which one acted,' that goes beyond a mere difference in form.

Secondly, and more generally, the kind of reason picked out by the 'reason why' expression is different to those picked out by all of the reason expressions already considered because its relata are different. While all the reasons picked out by the reason expressions already considered relate to actions, a *reason why* is related to a fact (or proposition). For instance, compare the relata of these expressions: 'my reason for swimming'; and 'the reason why I swam' – while swimming is an action, 'I swam' is a sentence.

The difference in relata is also of some relevance to our third observation: because reasons why relate to facts, the sorts of things they are reasons for need not be in any way agent involving. For instance, there are *reasons why* the Earth orbits the Sun and there are *reasons why* the Dawlish sea wall collapsed in 2014. Neither of these is an agent-involving occurrence; the Earth and the Dawlish sea wall aren't/weren't agents and they didn't do what they 'did' for reasons.

If we say that reasons that are essentially agent-involving are 'practical' kinds of reasons, then the kind of reasons picked out by the 'reason why' expression is not a practical kind of reason. This does not mean 'reasons why' never relate to agent-involving activities, that is, for instance, it does not mean that there can't be *reasons why* agents do things; it just means that 'reasons why' are not exclusively agent-involving (because they sometimes don't involve agents).

In contrast, there are no reasons for the Earth to orbit the Sun (or reasons for the Earth's orbiting the Sun, or reasons the Earth has to orbit the Sun), and there is nothing that could be called 'the Earth's reason for orbiting the Sun' because the kinds of reason picked out by these expressions are *practical*; so they don't apply to things that aren't agents, like the Earth. Unlike the 'reason why' expression, then, all the reason expressions of the previous sections pick out practical kinds of reason.[16]

So, the kind of reason picked out by the 'reason why' expression is distinct from the kinds of reason picked out by the expressions already considered (because it has different relata and it isn't *practical*). This isn't to say that something that is a reason of a practical kind can never

[16] Of course these remarks do not amount to a proof that the kinds of reason picked out by the previous expressions are practical – but I take that to be clear enough that we can assume it without controversy.

also be a reason why someone does something[17]; it is just to acknowledge that the conditions for being a practical reason differ from the conditions for being a reason why. That is, the conditions for being a reason of the kind picked out by the 'reason why' expression differ from the conditions for being a reason of the kind picked out by any of the other expressions; the 'reason why' expression picks out a different kind of reason to the reason expressions already considered.

## 1.7 A summary of reason expressions

I said that two kinds of reason are different kinds of reason if, and only if, the conditions for being a reason of one kind differ from being a reason of the other kind.[18] The discussion of the previous sections has established that there are *at least* three kinds of reason: independent & practical reasons; dependent & practical reasons; and independent & non-practical reasons (see Table I-2).

| **Reason expression** | **Dependent/Independent** | **Practical/Non-practical** |
|---|---|---|
| Reasons there are to act | Independent | Practical |
| Reasons for acting | Independent | Practical |
| Reasons one has to act | Independent | Practical |
| One's reason for acting | Dependent | Practical |
| Reasons why | Dependent | Non-practical |

**Table I-2:** Reason expressions and the kinds of reason they pick out

In addition, I have argued that, in spite of the fact that the standard categorisation schema ordinarily draws no distinction between the reasons there are to act, reasons for acting, and the reasons one has to act (that is, the expressions that pick out *independent* kinds of reason) or an agent's reason for acting and the reasons why they act (that is, the expressions that pick out *dependent* kinds of reason) we should nonetheless use a categorisation schema that allows for the *possibility* of distinguishing between them, in particular, so that we can discriminate between already existing theories.[19]

[17] Indeed, it is common to assume that reasons of the kind picked out by the expression 'the agent's reason for acting' are always reasons why the agent acts (see § (IV)1.2 for further discussion of this point).

[18] To stress a point I have already made: this definition is not meant to preclude the possibility of one and the same thing being a reason of two different kinds.

[19] My arguments were, in brief, that the difference in form provides a *prima facie* basis for believing that they pick out different kinds of reason, that some theories already distinguish between the reasons

Thus, I propose the following, provisional categorisation schema:

| | **Reasons there are to act** | **Reasons for acting** | **Reasons one has to act** | **Reasons for which one acts** | **Reasons why one acts** |
|---|---|---|---|---|---|
| *Theory of reasons* | *Claim…* | *Claim…* | *Claim…* | *Claim…* | *Claim…* |

**Table I-3:** A provisional categorisation schema

In the next section I will consider the character of the claims that differentiate theories of reasons. Having done so, I will suggest a revision to this provisional schema that omits the non-practical kind of reason, *reasons why*.

## 2 Claims about reasons

What I have called the 'standard schema' categorises theories on the basis of whether or not they take reasons of a given kind to be exclusively psychological. In doing so, however, this schema obscures the many differences between distinct anti-psychological theories. Indeed, what the standard schema calls 'anti-psychologism' is not so much a theory of reasons as it is a collection of different theories of reasons that, despite sharing a commitment to the falsity of the psychological view of any kind of reason, vary in many other respects.

For instance, different versions of anti-psychologism disagree on whether or not reasons can be false, and whether or not the existence of a reason depends on an agent's perspective – but the standard schema is incapable of recognising such disagreement. That is, of course, perfectly fine so long as one's focus is on psychologism and its discontents, but it becomes quite inappropriate when one wants to categorise theories of reasons from a more general perspective.

Furthermore, although there are alternatives to the psychological vs. anti-psychological categorisation that *can* differentiate between different forms of anti-psychologism, they, in contrast, obscure the differences between psychological theories and the different variants of anti-psychologism. For instance, Turri (2009) categorises theories of reasons according to their ontology – separating theories that take reasons to be facts, from those that take them to be mental states and again from those that take them to be intentional objects (that is, the contents of mental states). However, a strict[20] ontology-based categorisation is no less

there are to act and the reasons one has to act, and that I wish to distinguish between the reasons there are to act and reasons for acting.

[20] I say 'strict' because many take 'factualism' to be the view that reasons are facts *and* that they are not exclusively psychological (indeed, it's common to use 'factualism' to name the view that reasons are facts that favour actions that are consequently typically non-psychological). This, of course, is not a

problematic than the standard schema's approach, because (as Alvarez (2016b, 3) notes) it cannot discriminate between a psychological theory that says that reasons are all facts (about an agent's mind) and an anti-psychological theory that similarly says that reasons are all facts (about all kinds of things), since both theories have a common, factualist ontology.[21]

Alternatively, Lord (2015), Alvarez (2016a) and Way and Whiting (2017) divide theories according to whether or not they take the existence of a given kind of reason to depend on an agent's perspective. However, the categorisation of theories into 'perspectivist' and 'objectivist' similarly fails to distinguish perspectivist anti-psychological theories from (equally perspectivist) psychological theories.

One could, perhaps, surmount these problems by categorising the different theories on several or all of these aspects,[22] but I think that such a response would be to miss the point. The problem with all of these approaches is not that they each only capture one aspect of variation between theories. Rather, I suggest, it is that they put the cart before the horse in the sense that they classify theories of reasons according to their stance on a particular topic, rather than what determines that stance, which is, I suggest, what the theory actually is about.

To wit, the answers to the questions of whether or not reasons are psychological, of what their ontology is and of whether or not their existence depends on an agent's perspective are all determined by the answer to this, more fundamental question: for a given kind of reason, what is the relation between reasons of that kind and the actions for which they are reasons?

Let us call this relation the 'reason-relation': if you think that the *reason-relation* for a given kind of reason is such that it can relate non-psychological things to actions, then your theory will come out *anti-psychological* for that kind of reason, or if you think that the reason-relation for a given kind of reason is such that it is not a factive relation, then you will, perhaps, conclude that reasons of that kind are not facts (perhaps they are intentional objects – but that will depend on what you think the reason-relation is…). In this way, what a theory takes the reason-relation to be just determines its answer to these other questions.

The job of a theory of reasons is, *inter alia*, to say how many kinds of reason there are and to then explain what the reason-relation is for each kind of reason. The standard approach to

---

*strictly* ontological classification – an ontological classification is one that classifies theories on the basis of their ontology *alone*.

[21] The question of what they are facts *about* is not an ontological one, and so cannot feature in a strictly ontological classification.

[22] For instance, one could differentiate between theories that are perspectivist & psychological; perspectivist & anti-psychological; and objectivist & anti-psychological.

categorising theories puts the cart before the horse because rather than categorising them according to their claims about the reason-relation (which is the substance of the theories), it categorises them according to the *consequences* of their claims about the reason-relation. Of course, this may well be fine for the purposes to which the categorisation is habitually put (e.g. discussing the merits of psychological vs. anti-psychological or perspectivist vs. objectivist theories), but it is not fine for the purpose of actually categorising the different theories of reasons.[23] To do that, we should put the horse *before* the cart and categorise theories of reasons according to what they say about the reason-relation for each reason expression.

## 2.1 Families of claims

So what do they say? For any given reason expression, most theories of reasons subscribe to a view that belongs to one of three families: 'favourism', 'deliberativism' and 'psychologism'.[24] These three families do not exhaust the available claims about any given reason expression,[25] however, they cover, between them, the vast majority of claims made about each reason expression (both considered and *de facto*), and they are the focus of my discussion in what follows.

What the members of each of these families have in common is a view of what it takes to be a reason of any given kind: claims in the favourism family require that a reason of any given kind must make an action, in some respect, worth doing;[26] claims in the deliberativism family require that a reason of any given kind must be something the agent took to make their action, in some respect, worth doing; and claims in the psychologism family require that a reason of any given kind must be a feature of the agent's psychology that rationalises their action. Table I-4 sets out the claims for each reason expression in each family (at this stage I will leave the

[23] Nor, indeed, is it fine for really getting to grips with everything that is wrong with a particular theory – since it means concentrating on just one aspect of the theory.

[24] The name 'psychologism' I take from Dancy (2000), the others are my own invention.

[25] A noteworthy omission is Kearns and Star's (2008, 2009) account according to which reasons are evidence that one ought to act in a certain way.

[26] I have taken some liberties in representing a view that is typically rendered as 'reasons count in favour of actions' as 'reasons make actions worth doing'. I do so because leaving what it is to 'count in favour' of an action un-interpreted leaves the theory of reasons under-determined. I want 'favourism' to reflect the widely held interpretation of the 'counting in favour of' relation that takes it to either be the relation of 'making, in some respect, worth doing' or at least entailing that the action is, in some respect, worth doing. I specify favourist claims in terms of 'making, in some respect, worth doing' rather than 'favouring' because different interpretations of the 'counting in favour of' relation abound – for instance, Kearns and Star (2008) interpret what it is to 'count in favour' of acting as 'being evidence that one ought to so act', which is clearly different from what makes an action worth doing. Meanwhile Mitova (2016) advocates an alternative, if unspecified, construal of what it is to 'count in favour' of action, that is certainly not the idea of 'making it worth doing'. I return to this in § 4.3.

right-hand-side conditions of each account unexplained; they will be explained in subsequent chapters).

| | Family | | |
|---|---|---|---|
| **Reason expression** | **Favourism** | **Deliberativism** | **Psychologism** |
| For any *p, p* is a reason for *A* to *φ*… | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. | …if and only if *A* takes *p* to make *A*'s *φ*ing, in some respect, worth doing. | …if and only if *p* is a feature of *A's* psychology that rationalises *φ*ing. |
| For any *p, p* is a reason for *A's φ*ing… | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. | …if and only if *A* takes *p* to make *A*'s *φ*ing, in some respect, worth doing. | …if and only if *p* is a feature of *A's* psychology that rationalises *φ*ing. |
| For any *p, p* is a reason *A* has to *φ*… | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. | …if and only if *A* takes *p* to make *A*'s *φ*ing, in some respect, worth doing. | …if and only if *p* is a feature of *A's* psychology that rationalises *φ*ing. |
| For any *p, p* is *A*'s reason for *φ*ing… | …if and only if *p* makes *A*'s *φ*ing, *all things considered*, worth doing and explains (in the right way) why *A φ*'d.[27] | …if and only if *p* is a consideration in light of which *A φ*s. | …if and only if *p* is a feature of *A's* psychology that rationalises *φ*ing and explains (in the right way) why *A φ*'d. |
| For any *p, p* is a reason why *A φ*'d… | …if and only if *p* explains why *A φ*'d. | …if and only if *p* explains why *A φ*'d. | …if and only if *p* explains why *A φ*'d. |

**Table I-4:** The main claims about each reason expression

Favourism, deliberativism and psychologism are, as I have said, *families of claims* about reasons; they are not theories of reasons. Instead, a theory of reasons can be constructed by selecting one claim from one of the three families for each reason expression.

Within each family the conditions for being 'a reason there is to act'; 'a reason for acting' and 'a reason one has to act' are the same[28] – so a theory that holds the same view for any two of those expressions (say, *favourism about reasons there are to act* and *favourism about reasons one has to act*) takes them to pick out the same kind of reason. But, of course, the point of separating out the different reason expressions in our categorisation schema is that a theory *need not* always select a claim from the same family for each reason expression. That is, one could hold (as some do) *favourism about the reasons there are to act* and *deliberativism about the reasons one has to act* – and to do so is to say that those expressions pick out different kinds of reason.

[27] It's worth noting that a part of what is implicitly required for some *p* that makes an agent's action worth doing to explain their action *in the right way* is for it to be a consideration in light of which they act.

[28] They differ for 'the agent's reason for acting', as we noted they should, because we already know that that expression picks out a different kind of reason.

## 2.2 Revised categorisation schema

Finally, note that in Table I-4 there is no disagreement between different families about the conditions for being a *reason why*. As I have noted, what sets the 'reason why' expression apart from the other reason expressions considered is that it doesn't pick out a practical kind of reason. Instead, as the claims in Table I-4 make clear: *reasons why* are just explanations (of why someone does something, of why something is a certain way, of why something occurred, etc.).[29] They may be causal explanations (such as the reasons why the Earth orbits the Sun or the reasons why the Dawlish sea wall collapsed) but they equally may not be: the reasons *why* it is wrong to torture animals seemingly do not *cause* it to be wrong and the reasons *why* a football player is offside seemingly do not *cause* her to be offside – but they are reasons *why* it is wrong or reasons *why* she is offside all the same. Simply put: something is a reason why some other thing is the case if and only if it explains (causally or otherwise) why that other thing is the case.[30]

Of course, there is scope for disagreement about what the proper account of explanation is, and, acknowledging that, we could further analyse the 'explains' relation itself – this would create the possibility of more discriminating categorisations. However, I do not think we should seek to categorise theories of reasons on the basis of their preferred accounts of explanation – that is just a separate subject.

None of this is to say that explanation is not relevant to the discussion of practical reasons; indeed, some defend accounts according to which reason-relations just are a particular sort of explanatory relation.[31] My point is only that we can omit the 'reasons why' expression from our categorisation schema, since whatever dispute there is about it is properly a part of a separate discussion.[32]

[29] See, for instance, Raz (2009) and Alvarez (2010) for this view.
[30] On this analysis, '*p* is a reason why *q*' and '*p* explains why *q*' and '*q* because *p*' are logically equivalent.
[31] For instance, Broome (2006) argues that a reason to act is a reason why someone ought to do something (see § (II)3 for further discussion of this). And indeed, I will eventually defend the view that practical reasons are all reasons why an action is *pro tanto* rational.
[32] As Raz notes: 'Whatever one can say about [reasons why] is better explored when studying explanations, a voluminous philosophical subject.' (2009, 186)

Those remarks having been made, the revised framework for the schema I am proposing is, thus, as follows:

| | **Reasons there are to act** | **Reasons for acting** | **Reasons one has to act** | **Reasons for which one acts** |
|---|---|---|---|---|
| *Theory of reasons* | *Claim…* | *Claim…* | *Claim…* | *Claim…* |

**Table I-5:** A revised categorisation schema

# 3 Categorising theories of reasons

For the mathematically inclined: within my categorisation schema, a theory of reasons can be specified as a 4-tuple, with each member of the tuple being a claim about a reason expression from a particular family, ordered as follows: 'reasons there are to act'; 'reasons for acting'; 'reasons one has to act'; 'reasons for which one acts'. If we let F, D, and P denote Favourism, Deliberativism and Psychologism about the relevant reason expression, respectively, then, for instance, we can specify some distinct theories of reasons as follows: (F, F, F, F); (P, P, P, P); (F, F, D, D); (F, F, D, F). By way of interpretation: the first and second of these theories make no distinction between any *independent* kinds of reason, while the third and fourth distinguish between 'the reasons there are to act' and 'the reasons one has to act'.

I have set out the reason expressions that are to be included in our categorisation schema and I have set out the main claims that different theories of reasons make about each reason expression, as well as a (formal) way of describing theories of reasons in terms of these claims. Using this framework, in Table I-6, I categorise different theories of reasons.

This categorisation is far from exhaustive.[33] Moreover, the theories of some of the theorists that I have grouped together differ in some of their intricacies; more discriminating characterisations of the different *families* of claims could, perhaps, separate out those intricacies – but that level of discrimination does not alter the main criticisms that the theories are subject to.[34] Since I only need the categorisation to be fine-grained enough to sort between what arguments apply to which theories, this will do, and, in particular, it supports that end far better than the standard schema.

[33] Noteworthy omissions include John Broome and Thomas Scanlon, whom I struggled to place. I have also omitted Michael Smith from this categorization, whose work is typically associated with 'The Received View', because, as I will argue in § (V)4.1, he advocates a pluralist theory of reasons, which evades categorisation within this schema.

[34] See § (II)3.3 for more on this.

| Theories of reasons | *4-tuple description* | Reasons there are to act | Reasons for acting | Reasons one has to act | Reasons for which one acts |
|---|---|---|---|---|---|
| 'The Received View' | (F, F, F, P) | Favourism | Favourism | Favourism | Psychologism |
| Stout (2009), Alvarez (2010), Parfit (2011), Littlejohn (2012) | (F, F, F, F) | Favourism | Favourism | Favourism | Favourism |
| Dancy (2000, 2014), Davis (2005), Sandis (2009) | (F, F, F, D) | Favourism | Favourism | Favourism | Deliberativism |
| Schroeder (2008), Comesaña & McGrath (2014) | (F, F, D, D) | Favourism | Favourism | Deliberativism | Deliberativism |
| Hornsby (2008) | (F, F, D, D) | Favourism | Favourism | Deliberativism | Favourism |
| Turri (2009), Gibbons (2010), Mitova (2015) | (P, P, P, P) | Psychologism | Psychologism | Psychologism | Psychologism |

**Table I-6:** Prominent theories of reasons, categorised in my proposed schema

# 4 Normative and motivating reasons

The terminology of 'normative' and 'motivating' reasons is common throughout the literature on practical reasons, and, as I have noted, the standard schema typically categorises theories according to what they say about each. We are now in a position to see what is wrong with this terminology.

As I will argue, common definitions of these terms covertly import a theory of reasons into one's categorisation schema that leaves little scope for disagreement. Indeed, I argue that, from a rhetorical standpoint, we ought to see these definitions less as terminological housekeeping and more as a Trojan horse left by the anti-psychologists for the psychologists. So, I suggest, when working towards a theory of reasons in earnest, we should abandon this terminology, at least as it is standardly defined.

## 4.1 Terminology or Trojan horse?

The *de facto* distinction between *normative* (or *justifying*) reasons and *motivating* (or *explanatory*) reasons is probably as follows: *normative* reasons explain why an agent *ought* to do something whereas *motivating* reasons explain why they did it. While this distinction is

itself problematic,[35] my focus in this section is on the following, increasingly typical characterisations of these terms[36]:

- For any *p, p* is a normative reason for *A* to *φ* if and only if *p* counts in favour of *A*'s *φ*ing.
- For any *p, p* is *A*'s motivating reason for *φ*ing only if *p* is something *A* took to count in favour of *φ*ing.[37]

These characterisations are innocuous as they are: one can define one's terms however one wishes. However, what makes their use in theorising problematic is that their definitions aren't typically restricted to these innocuous statements, but also include the following claims[38]:

- For any *p, p* is a normative reason for *A* to *φ* if and only if *p* is a reason for *A* to *φ*.
- For any *p, p* is a normative reason for *A* to *φ* if and only if *p* is a reason for *A*'s *φ*ing.
- For any *p, p* is a normative reason for *A* to *φ* if and only if *p* is a reason *A* has to *φ*.
- For any *p, p* is *A*'s motivating reason for *φ*ing if and only if *p* is a reason for which *A φ*'d.

I call these latter remarks 'claims' because, by associating these terms *also* with reason expressions, these additional remarks go beyond mere definitions and well into what it is properly the business of a theory of reasons to determine.

[35] For instance: not everything that explains an agent's action is the sort of thing we would want to call a 'motivating reason' (see Alvarez 2010).

[36] For instance: 'Most contemporary philosophers start by distinguishing two types of reason for action: "normative" reasons – that is, reasons which, very roughly, favour or justify an action, as judged by a well-informed, impartial observer; and "motivating" reasons – which, again roughly, are reasons the "agent" (that is, the person acting) takes to favour and justify her action and that guides her in acting.' (Alvarez 2016a, 1) It's worth noting that this is not how motivating reasons were always defined, in particular, Michael Smith (1987) has quite a different notion in mind, as Darwall (2003) observes – see § (V)4 for further discussion of this point.

[37] Note that this is only a necessary condition, not a sufficient one. As I have noted, it is common to assume that motivating reasons (so defined) also explain an agent's action – although not everyone takes that view (e.g. Davis 2005).

[38] For instance: 'There are normative reasons: *reasons that there are for people to act* – as it is often put, reasons that 'favour' doing something; and motivating reasons: *reasons for which an agent acts*, that is, the reasons that an agent takes (perhaps rightly) to favour acting as she does and for which she acts.' (Alvarez 2016b, 4–5 emphasis added) Similarly: 'Normative reasons are considerations which count in favour of, or against, an action. What you ought to do is determined by how the normative reasons for and against acting weigh up – roughly, you ought to do what the balance of such reasons supports. For instance, if *there is a reason for you to take an umbrella* [i.e. a normative reason to take an umbrella], and no stronger reason not to do so, you ought to take an umbrella. Normative reasons contrast with motivating reasons – *the reasons for which you act*. In some cases, the reasons for which you act are, or correspond to, *reasons for acting*. That is to say, in some cases, your motivating reasons are, or correspond to, normative reasons. For instance, that it is raining might speak in favour of taking an umbrella and be the reason for which you do so.' (Way and Whiting 2017, 2–3 emphasis added)

In particular, as well as precluding the possibility of distinguishing between the kinds of reason picked out by the expressions 'the reasons there are to act', 'the reasons for acting' and 'the reasons one has to act' (the issues with which I have already discussed at length), these 'definitions' go some way to setting out what the reason-relation for each reason expression is, which is to say that they go some way to setting out a theory of reasons. That is, behind the standard schema's seemingly innocuous definitions of 'normative' and 'motivating' reasons is the 'assumed theory' of reasons set out in Table I-7.

The problem is that if we *start* our theorising with the 'assumed theory' in mind then all that is really left to debate is how to interpret the relation of 'counting in favour of', which, as we shall see, gives anti-psychological theories a significant advantage. And it is precisely this advantage that makes me suggest that this act of seeming terminological housekeeping ought to be seen as a Trojan horse, as the next section discusses.

| **Reason expression** | **The assumed theory** |
|---|---|
| For any *p, p* is a reason for *A* to *φ*… | …if and only if *p* counts in favour of *φ*ing. |
| For any *p, p* is a reason for *A's φ*ing… | …if and only if *p* counts in favour of *φ*ing. |
| For any *p, p* is a reason *A* has to *φ*… | …if and only if *p* counts in favour of *φ*ing. |
| For any *p, p* is a *A*'s reason for *φ*ing… | …only if *A* took *p* to count in favour *φ*ing.[39] |

**Table I-7:** The 'assumed theory' of reasons

## 4.2 Why it is a Trojan horse

When it is raining and I believe that it is raining, what is it that counts in favour of taking an umbrella – is it the fact that it is raining or the fact that I believe that it is raining? The most natural interpretation seems to be that, in ordinary circumstances, it is the fact that it is raining that counts in favour of taking my umbrella, and not the fact that I believe that it is raining.

A classic[40] sort of example illustrates this well: suppose that Sam believes that the security services are trying to read her mind. If it were true that the security services were trying to read her mind, then the fact that they were trying to read her mind would count in favour of her wearing a foil hat.[41] However, the fact that she *believes* that the security services are trying

[39] It's worth noting that because this doesn't provide sufficient conditions for being an agent's reason for acting, the assumed theory is not a complete theory of reasons – there is, as I will later note, still some room for disagreement, albeit modest.

[40] Alvarez (2016b) attributes this example to Anscombe (1957); Hyman (1999); Raz (1999b) and Dancy (2000).

[41] Because, *inter alia*, foil hats block the radiofrequency electromagnetic radiation that the security services use to read minds (or would, if they did).

to read her mind does not favour wearing a foil hat – it favours going to see a doctor.[42] What this example shows is that the circumstances in which a psychological fact 'counts in favour' of an action *in the same way* as the (non-psychological) fact that it is raining 'counts in favour' of taking an umbrella are unusual – it is typically features of the world that count in favour of doing things, not features of our psychology. That is, if a well-informed observer were to list the things that count in favour or count against some action, they would rarely list facts about the agent's beliefs as things that count in favour (or against) their action (however much those facts are likely to affect what the agent actually does).

We can press the point about the interpretation of the 'counting in favour of' relation by considering what it is that an agent *takes* to count in favour of their actions.[43] Seemingly, what Sam takes to count in favour of wearing a foil hat is *that the security services are trying to read her mind* (and this is so even if they aren't), in contrast, what she takes to count in favour of going to see the doctor is *that she believes that the security services are trying to read her mind*. Sam is deliberating both about the way she takes the world to be and *the fact that she takes it to be that way*. Again, what the story about Sam's deliberation shows is the unusualness of deliberating about features of one's own psychology – we don't normally take features of our psychology to favour a given course of action – it is normally only the *contents* of our mental states that feature in our deliberation.

This line of reasoning leads us rather quickly from 'the assumed theory' to the following conclusions: firstly, since the things that count in favour of actions (and are taken to count in favour of actions) are usually *not* psychological, all the claims in the psychologism family must be false (since they only allow for psychological reasons). So whatever our theory of reasons is, it must be anti-psychological with respect to *every* reason expression.

Secondly, since something only counts in favour of an action if the action is, in some respect, worth doing (the fact that it's raining doesn't count in favour of taking my umbrella if taking my umbrella is, *to no extent*, worth doing (if I actually *want* to get wet, say)), and since it is seemingly worth doing in virtue of that which counts in favour of it, the natural interpretation of what it is to 'count in favour of' an action seems to be this:

[42] Because, in Sam's case, her belief that the security services are trying to read her mind is the product of a delusional disorder.

[43] Noting that (which, as Table I-7 makes clear) being something that the agent takes to count in favour of an action is a necessary condition on being the reason for which an agent does that action.

- For any *p, p* counts in favour of *A*'s *φ*ing if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing.[44]

By now we should smell a rat; substituting the above, which I have suggested is the most natural interpretation of the 'counting in favour of' relation, into Table I-7, yields the following theory:

| **Reason expression** | **The assumed theory – conventionally interpreted** |
|---|---|
| For any *p, p* is a reason for *A* to *φ*… | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p, p* is a reason for *A's φ*ing… | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p, p* is a reason *A* has to *φ*… | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p, p* is *A*'s reason for *φ*ing… | …only if *A* took *p* to make *A*'s *φ*ing, in some respect, worth doing. |

**Table I-8:** The conventional interpretation of the assumed theory

The theory set out in Table I-8 should look familiar; the claims of the re-interpreted assumed theory for the first three reason expressions just are the claims of *favourism* about those reason expressions.

So, my concern is this: the terminology of 'normative' and 'motivating' reasons amounts to a Trojan horse because once you accept it, the most natural interpretation of the 'counting in favour of' relation then fixes your theory for the first three reason expressions (that is, it forces you to accept *favourism* about those reason expressions). Having accepted the assumed theory all that is left to debate is then whether to endorse favourism or deliberativism about the reasons for which we act (which are both consistent with the final row of Table I-8[45]). That is, what makes the terminology a Trojan horse is that it forces many theorists[46] to accept an account of the reason-relation for each reason expression that puts them at a systematic disadvantage – the only way to advocate their theory is for them to adopt an unnatural interpretation of what it is to 'count in favour of' an action.

## 4.3 The Trojan horse rejected

The Trojans ought to have left that damned horse alone, and so too should we, if we are to do our theorising in earnest. That is, rather than accommodating different theories by adopting

[44] I say this without meaning to undermine the view that the relation of 'counting in favour of' is somehow primitive. It is, I think, still possible that the favouring relation is the more fundamental one even if this is the case – see § (II)3.1 for further discussion of this point.

[45] As fn. 39 records – the final row omits a sufficient condition for being a reason for which the agent acts, so the assumed theory falls short of a full theory of reasons.

[46] That is, theories of reasons other than (F, F, F, F) or (F, F, F, D).

an unnatural interpretation of what it is to 'count in favour of an action', [47] the best response is to reject the horse by refusing to accept the assumed theory as the starting point of theorising.

And once one does that, the oft-repeated argument against psychologistic theories[48] becomes question-begging: it starts off by assuming the truth of a position[49] that psychologistic theories should reject.[50]

## 4.4 Is it a Trojan horse or is it just the truth?

But is it really a Trojan horse? Isn't counting in favour of acting just *what it is* to be a reason? This is, after all, a well-established view, brought to the fore in the following remarks:

> I will take the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. 'Counts in favor how?' one might ask. 'By providing a reason for it' seems to be the only answer. (Scanlon 1998, 17)

> If we are asked what reasons are, it is hard to give a helpful answer. Facts give us reasons, we might say, when they count in favour of our having some belief or desire, or acting in some way. But 'counts in favour of' means 'is a reason for'. Like some other fundamental concepts, such as those of reality, necessity, and time, the concept of a reason cannot be explained in other terms. (Parfit 2001, 18)

The suggestion is that 'being a reason for' and 'counting in favour of' are interchangeable expressions. Now, were that view immediately obvious then what I have called a 'Trojan horse' would be less of a covert assault on psychologistic theories and more just the inconvenient truth for them. I don't, however, think that this view is immediately obvious, particularly because I think it is actually false.

While I agree with Scanlon and Parfit that whenever something is a reason for acting it must count in favour of so acting, [51] I disagree with them in so far as I think that it counts in favour so acting as a consequence of its being a reason *for* acting, and not a consequence of its being a reason. In particular, I do not think it follows that anything that counts in favour of an action is a reason for doing it, nor that all reasons count in favour of actions.[52] What I want to argue

---

[47] According to which the fact that Sam believes that the security services are trying to read her mind *counts in favour* of wearing a tin foil hat. I think that Mitova (2016) prefers that option.

[48] By which I mean the argument set out in § 4.2, which one could put succinctly as follows: what counts in favour of actions is often not psychological (as shown by Sam's case) and that which counts in favour of an action is a reason to do it, so reasons can be things other than features of an agent's psychology, so psychologism (which says that they can't) is false.

[49] I.e. that something is a reason just in case it counts in favour of an action.

[50] This is, I think, precisely Gibbons's (2010, 354) objection to the use of this line of reasoning against his 'psychologistic' theory.

[51] And that whenever something is a reason *against* acting it must count against doing it.

[52] That is, while I agree that if *p* is a reason *for φ*ing then *p* counts in favour of *φ*ing. I don't agree that the right-to-left reading (if *p* counts in favour of *φ*ing then *p* is a reason for *φ*ing) is true, nor do I agree that if *p* is a reason for *A to φ* then *p* counts in favour of *φ*ing.

for now, however, is not the view that 'being a reason' and 'counting in favour of' aren't interchangeable expressions – but just that the view that they are is not *immediately obvious*. My argument is this: it seems more likely that the meaning of 'counting in favour of' comes from the fact that something is a reason for acting than that it is a reason for acting.

The *Oxford English Dictionary* defines the 'for' preposition (*inter alia*) as follows:

> 'In defence or support of; *in favour of*, on the side of. Opposed to against.' ('For, Prep. and Conj.', n.d., 7a emphasis added)

Seemingly anything that is *for* something, in this sense of the preposition, counts in favour of it, in virtue of its being *for* it (and not necessarily anything else about it). So, I suggest, what makes a 'reason for acting' count in favour of an action is the fact that it is *for* that action (as opposed to against it), and not the fact that it is a reason. That is, my claim is that we should not *necessarily* analyse 'being a reason' as 'counting in favour of', rather we have stronger reasons for analysing 'being *for* something' as 'counting in favour of it'.

This is made clearer by the fact that the 'for' preposition lends the meaning of 'counts in favour of' to things other than reasons. For instance, I can say: 'the factors for and against acting'; 'the things for and against acting'; 'the considerations for and against acting'; 'what there is to be said for and against acting'. Wherever this 'for' appears, we can say that the thing that precedes it counts in favour of that which it is *for*, but that doesn't lead us to say that what it is to be a 'factor', or a 'thing', or a 'consideration', or 'something to be said', depends on counting in favour of some action independent of its being a factor *for*, or a thing *for*, or a consideration *for*, or something to be said *for* doing an action. Likewise, I suggest, the fact that being a reason *for* doing something depends on counting in favour of doing it, should not lead us to assume that being a *reason* depends on counting in favour of doing something – instead we should just admit that whenever anything is *for* something, it counts in favour of it, and just because it is *for* it. [53]

To be clear: my aim with these remarks is not to show that some reasons don't count in favour of acting. My aim is just to show that it is not a foregone conclusion that all reasons count in favour of actions, and it is certainly not something that we should assume at the outset of theorising. One's theory might end up arguing that the reason-relations are best understood in terms of the 'counting in favour of' relation (i.e. as favourists do) – but that is an argument that must be made, it is not the default position. That is, we should not *start off* our theorising

[53] What about the agent's reason *for* acting? Doesn't that then count in favour of their action, since it is *for* it? I do not think it does, I suggest that this use of 'for' appeals to one of the prepositions (many) other meanings; specifically 'of purpose or destination.' ('For, Prep. and Conj.', n.d., IV) The association of this expression with this meaning is discussed in more detail in § (II)1.2.

by assuming either that something is a reason to act, a reason for acting or a reason one has to act just in case it counts in favour of acting, or that a reason for which an agent acts is always something they take to count in favour of their action. To start our theorising in this manner is, I submit, prejudicial to our enquiry. And this, accordingly, is why I have rejected the standard schema and why I eschew the terminology of 'normative' and 'motivating' reasons.

# 5 Conclusion

I have argued that the standard schema is ill-suited to the task of categorising different theories of reasons and I have proposed an alternative categorisation. My alternative categorisation distinguishes theories according to what they say the reason-relation is for a number of typical reason expressions and it eschews the language of 'normative' and 'motivating' reasons because those terms (as they are standardly defined) already involve substantive theoretical commitments.

In the next three chapters I discuss the main critiques of what are typically taken to be the strongest suits of each of the three families, namely (and I discuss them in this order): favourism about reasons to act; psychologism about the reasons for which we act; and deliberativism about the reasons for which we act. I set out a number of *prima facie* reasonable claims and show how each of these views must reject some subset of these claims.

# (II)

# Reasons to act that make actions worth doing

*In which I show what it costs to think that if there is a reason to do some action then that action is, in some respect, worth doing. I show how 'favourism about reasons to act' (which entails that reasons to act make actions worth doing) clashes with some* prima facie *reasonable claims about a case in which someone acts on a false belief. I set out which claims the proponent of this view must choose between rejecting and I categorise the common choices from the literature.*

While walking through a forest that she knows to contain bears, Sally hears what sounds like a bear running towards her.[1] She runs, frantically, to a nearby safe-house. In fact, no bear was chasing her; it was an odd rustling of the trees that made the noise. Did Sally have a reason to run? Of course – she thought a bear was chasing her! Did she run for a reason? Surely! What was her reason for running? She thought that a bear was chasing her, or perhaps one might say this: she heard a bear-like sound.

This seemingly straightforward story and these typical intuitions about it, create problems for what I have called 'favourism about reasons to act', according to which a reason for someone to do some action is a fact that makes it, in some respect, worth doing.

The purpose of this chapter is to set out a series of *prima facie* reasonable claims about reasons[2] and to then show that 'favourism about reasons to act' must reject at least some of them. This is not meant to be a conclusive argument against this view – only to show that it comes at the cost of rejecting some *prima facie* reasonable claims.

## 1 Sally and the non-existent bear

### 1.1 Reasons and rationality

Given that she knew that safety was nearby, and given that she thought that a bear was chasing her (add that she wants to live, if you like), it was plainly rational for Sally to run, as Stout notes:

> What can be accepted without much difficulty is that her having that belief makes her running away rationally intelligible. Learning that she thinks a bear is chasing her I can make sense of her running away; I can see that her behaviour is rational. (Stout 2009, 52)

[1] This example is adapted from Stout (2009).

[2] Based both on this story about Sally and on more general intuitions about how 'the reasons there are for an agent to act' relate to 'the reasons they have to act' and 'the reasons for which they act'.

Since it was plainly rational for Sally to run, let's say it plainly:

(F1) It was rational for Sally to run.[3]

Now, I said that Sally had a reason to run, even though no bear was chasing her. This seems like a natural thing to say, but why? I think that it is because, as Errol Lord observes:

> It's natural to think that whenever it's rational for me to *φ*, I have reasons to *φ*. (Lord 2010, 1)[4]

It's natural to say that she has a reason to run, and it sounds strange to say that she doesn't because the fact that it was *rational* for her to do it suggests that she *had a reason* to do it. So, here is another *prima facie* reasonable claim:

(F2) If it is rational for *A* to *φ* then some *p* was a reason *A* had to *φ*.

## 1.2 Acting for a reason and acting intentionally

I said that Sally ran for a reason. To deny this and say that Sally didn't run for a reason, that is, that she was running for no reason, seems to suggest that her action was unconsidered, that, at best, she was running for the sake of running, or running on a whim. If we were to say to someone that Sally ran for no reason, and then add that, by the way, she ran because she thought a bear was chasing her, it would seem to cancel the sense in which she ran for no reason: she is certainly not running for the sake of running if she is running because she thinks that a bear is chasing her.

The point is not that there is no way of making the claim that she ran for no reason intelligible – it's just that saying that she ran for no reason *and* that she ran because she thought a bear was chasing her has an air of contradiction about it. And the reason it has an air of contradiction about it is that saying that someone didn't act for a reason implies that their action lacked some sort of rational, intentional, deliberateness – it makes it seem whimsical or unconsidered, if not entirely unintentional. As Dancy notes:

> Intentional, deliberate, purposeful action is always done for a reason. (Dancy 2000, 1)

Why does Dancy choose the particular string of adjectives ('intentional, deliberate, purposeful') he does? I think he is trying to anticipate an objection to the simpler claim that *all* intentional action is done for a reason. In particular: one might say that if I cross my legs (to use Dancy's example), I act intentionally but I don't act for a reason – the sort of acts one does

[3] You could, of course, add in defeaters (she thought the bear was between her and the safe-house; she thinks that staying still is the best way to avoid a bear attack), but that would be to change my story – there are no such defeaters here; the most rational thing for Sally to do is run.

[4] For similar remarks on the 'naturalness' of this claim, from philosophers with different views about reasons see Unger (1978, 200), Alvarez (2010, 13), Gibbons (2010, 337) and Comesaña and McGrath (2014, 61).

on a whim may be done for no reason, but they aren't unintentional. Dancy, quite rightly, doesn't want to us to think of such actions (which O'Shaughnessy (1980) calls 'sub-intentional' actions), as actions that are done intentionally, purposefully and deliberately.[5] It seems that there is a clear class of what we might call *sophisticatedly* deliberate actions that we associate with acting for a reason.[6] For the sake of brevity I will shorten Dancy's string of adjectives to just *deliberate* action and formulate this *prima facie* reasonable claim as follows:

(F3) If *A φs* deliberately then *A φs* for a reason.

This view is shared by philosophers with a wide range of views on what reasons are,[7] and, indeed, Mele remarks that 'the overwhelming majority of ordinary speakers of English asked for a gut reaction to [this claim] would find it extremely plausible.' (Mele 2007, 99)

Did Sally run deliberately (intentionally, purposefully…)? Of course she did! She responded to the situation as she took it to be, she decided on the appropriate course of action and she acted on the decision she'd made (albeit hurriedly!).[8] Thus we can say:

(F4) Sally ran deliberately.

So this is why it sounds odd to say the Sally didn't run for a reason: it suggests that she didn't run *deliberately* (intentionally, purposefully…) – which she plainly did.

## 1.3 Sally's reasons for running

The remarks I made about Sally's reasons for running bear (!) repeating, since I take them to be *prima facie* reasonable. Firstly, I said that:

(F5) Sally's reason for running was, *inter alia*, that she believed a bear was chasing her.

[5] Mele (2007, 99) considers another would-be counterexample to claim that all intentional action is done for a reason, which Dancy's wording is seemingly also designed to avoid.

[6] It's worth noting that the sense of sophistication here is purely internal to the agent – what makes something count as, in Dancy's terminology, an intentional, deliberate and purposeful action is all about the way the agent reasoned their way to it, and nothing to do with its correspondence to the external world.

[7] For instance, it's worth noting that Maria Alvarez, who rejects (F3) nonetheless notes that 'The suggestion that someone who acted motivated by a false belief does not act for a reason might seem prima facie wrong.' (2010, 141) That is all that I am saying – that it is *prima facie* reasonable to say that they do act for a reason. (See also Anscombe 1957, 9; Davidson 2001c, 83; Davis 2005, 68–69; Gibbons 2010, 357; Hieronymi 2011, 410–11)

[8] Sally's circumstances are extreme but I don't think it's unreasonable to suggest that she deliberates about what to do. She might be torn, for instance, over whether it would be better to 'play dead' – can she out run the bear? One might well, I submit, deliberate even when the stakes are high and time is short – such deliberation need not take a long time, or even have much to it (the matter might be straightforward) for one's action to be deliberate.

It flies in the face of experience to deny that we make such remarks, as, indeed, Joseph Raz notes:

> There is no denying that we use locutions of the form 'his reason for *φ*ing was his belief that *p*'… (as in 'his reason for not coming was that he thought you would not be here'). (Raz 1999b, 18)

I also said that:

(F6) Sally's reason for running was, *inter alia*, that she heard a bear-like sound.

This seems to suggest that an appearance or perceptual experience could be an agent's reason for acting. I find it as natural to say this as saying that her reason for running was that she believed that a bear was chasing her, but it is perhaps worth noting that reasons of this kind are less commonly discussed in the literature.[9] Nonetheless, I take this claim to be, equally, *prima facie* reasonable.

## 2 How reason expressions relate

In § (I)1, I discussed different reason expressions, but made no mention of the relations between them. So how are they related? Here is what Gibbons suggests:

> If you go to the store for milk… you will go there for a reason. So you must have a reason. So there must be a reason. (Gibbons 2010, 343)

Gibbons' inference seems natural. That is, it seems *prima facie* reasonable to think that if one acts for a reason one must have had a reason to act and if one had a reason to act then there must have been a reason to act, which one had. So:

(F7) If *A φs* for a reason then there was a reason, *p*, for *A* to *φ*.

Denying this means, as Jennifer Hornsby notes (albeit, in the process of denying it), saying things like: 'there was no reason to do what he did, even though he did it for a reason,' (2008, 249) which is, in her own words, '*prima facie* paradoxical.' The denial of a *prima facie* paradoxical claim is, it seems to me, at least *prima facie* reasonable.

I think that Gibbons' inference permits of some strengthening: it is not just that you go to the store for a reason *and* you have a reason *and* there is a reason, but that your reason for going there *is* a reason that you have to go there and *is* a reason that there is for you to go there.

[9] Although Dancy alludes to this sort of reason for acting when he notes that when Edmund stays away from the middle of an icy lake after his friend had told him that the ice was thin there, 'his reason could have been simply that his friend had warned him off.' (Dancy 2014, 88) Kearns and Star (2008) defend (at length) an account that agrees with this intuition and Whiting (2014) also notes that we often cite perceptual experiences as reasons. Lord (2010, 6) is also clear that a reason to believe something is a reason to do that which the belief makes rational – assuming that experiences can be reasons for belief, he would seemingly agree with (F6).

These strengthened claims provide us with the following, additional *prima facie* reasonable claims about reasons:

(F8) For any *p,* if *p* was a reason *A* had to *φ* then *p* was a reason for *A* to *φ*.[10]

(F9) For any *p,* if *A φs* for the reason that *p* then *p* was a reason for *A* to *φ*.[11]

Both of these claims are at least *prima facie* reasonable because denying either amounts to saying, implausibly, that something that is a reason isn't a reason: denying (F8) means saying that an agent has a reason to act that isn't a reason to act; and denying (F9) means saying, as Dancy notes, 'an agent can act for a reason that is no reason.' (2000, 3)

Indeed, Mele finds it so strange to say such things that he makes it an explicit constraint on a satisfactory theory of reasons that it concede that: 'anything that is a reason is not no reason.' (2007, 95) I am not going as far as Mele – I note his remarks only to show that these claims enjoy considerable intuitive plausibility, so they are at least *prima facie* reasonable, and that if a theory rejects them, it is at some cost that it does so.

# 3 Favourism about reasons to act

Recall the following claim about reasons from Table I-4:

> *Favourism about reasons to act*: For any *p, p* is a reason for *A* to *φ* if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing.

This is not the standard way of characterising the view of those to whom I attribute it.[12] These theorists more typically characterise their view as the claim that a reason to do some action is that which 'counts in favour' of doing it.

However, I characterise it in the way that I do because there are differences of opinion about what it is to 'count in favour' of an action,[13] so the problems for favourism about reasons to act (to be discussed in § 4) do not apply to everyone who holds the view that reasons to act count in favour of actions. My aim in characterising the view in this way is thus to group together those theorists (of which there are many) who are susceptible to the problems that

[10] Alvarez (2016b, 11) provides a thorough defence of the intuitiveness of *this* claim. It is also worth noting that I make no claim about the truth or falsity of the right-to-left reading here.

[11] Cf. Dancy's (2000) normative constraint, which is a weaker version of this thesis. It says that if an agent acts for the reason that *p* then *p could have been* a reason there was for them to act.

[12] See Table I-6 for the list of those to whom I attribute this view.

[13] For instance, Kearns and Star suggest that 'a fact counts in favor of *φ*ing just in case this fact is evidence that one ought to *φ*.' (2008, 44) And Mitova (2016) also rejects the conventional interpretation, although she leaves the work of developing an alternative for another time. I also think Lord (2010) requires another interpretation of the 'counting in favour of' relation if he is to get his desired result that that which justifies a belief also counts in favour of an action that the belief makes rational.

follow, acknowledging that there are subtle variations between their theories (more on this to come).

In what follows I provide some context for how I arrive at the characterisation that I do, as well as some explication of what it is for an action to be, in some respect, worth doing.

## 3.1 'Counting in favour of'

As Gjelsvik (2007) notes, there are differing ways of conceiving of what it is to 'count in favour of' an action that arrive at seemingly the same conclusion about when an action is favoured. To understand the first way, let us revisit the following remarks:

> I will take the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. 'Counts in favor how?' one might ask. 'By providing a reason for it' seems to be the only answer. (Scanlon 1998, 17)

> If we are asked what reasons are, it is hard to give a helpful answer. Facts give us reasons, we might say, when they count in favour of our having some belief or desire, or acting in some way. But 'counts in favour of' means 'is a reason for'. Like some other fundamental concepts, such as those of reality, necessity, and time, the concept of a reason cannot be explained in other terms. (Parfit 2001, 18)

I first discussed these remarks in § (I)4.4, when I drew attention to the fact that Scanlon and Parfit take the 'counting in favour of' relation and the 'being a reason to' relation to be equivalent. This is something they have in common with, I think, everyone who endorses favourism about reasons to act. But another observation that we can make about these remarks, and one which, as we shall see, separates Scanlon and Parfit from some others to whom I attribute favourism about reasons to act, is that the former take the 'counting in favour of' relation to be primitive. This construal, ('favouring' as a primitive relation) is the first way of conceiving of what it is to count in favour of an action.

The main alternative is given its clearest exposition by John Broome, who states that 'to count in favour of $\varphi$ is to play a particular role in an explanation of why you ought to $\varphi$' (2006, 41). John Hyman (e.g. 2015, 133–34) also seems to be clear about the explanatory character of the reason-relation, and while he less directly analyses the 'counting in favour of' relation in those terms, I think his position amounts to the same.

Others are less explicit, but, I think, more or less implicitly analyse favouring as such an explanatory relation. For instance, some authors characterise 'counting in favour of' as making good, or right or valuable:

> A reason for action is something that favours or *makes valuable* an action of the relevant kind. (Everson 2009, 22 emphasis added)

> If a person has a reason for *φ*ing then it follows that that person ought to or may (if only *pro tanto*) *φ*. This might mean that the reason favours *φ*ing, or recommends, permits, warrants, or demands, etc. *φ*ing. But the question arises why a reason for *φ*ing favours, warrants, or demands *φ*ing? I suggested that the answer to this question is that *a reason for φing makes φing right or appropriate* (sometimes merely *pro tanto* right or appropriate). (Alvarez 2010, 12–13 emphasis added)
>
> When we think of such reasons, we think of features that speak in favour of the action (or against it)… they *make actions right or wrong, sensible or unwise*. (Dancy 2000, 1 emphasis added)

I think that it is natural to interpret the 'making' relations invoked here (reasons make actions valuable, right, sensible…) as explanatory relations. I find it odd to think that some *x* could make some *y* an *F* without (at least partly) explaining why it is an *F*, so I likewise find it odd to think that some reason could make some action right or valuable without (at least partly) explaining why it was right or valuable. Thus, I take accounts that say that reasons make actions right, valuable or good to be, albeit implicitly, of the same explanatory view as Broome.[14] If they aren't then there is perhaps a third way of arriving at what is ultimately the same conclusion.

What is that conclusion? I think we can put it,[15] as I did in hopefully un-contentiously, thus:

- For any *p, p* counts in favour of *A*'s *φ*ing if and only if *A*'s *φ*ing is, in some respect, worth doing.

Given the presumed equivalence of 'being a reason to' and 'counting in favour of' amongst advocates of favourism about reasons to act, we thus arrive at the view, set out above, that there is a reason for an agent to *φ* if, and only if, *φ*ing is, in some respect, 'worth doing' for that agent. But what does it mean to be for an act to be, in some respect, worth doing?

## 3.2 What it is for an act to be, in some respect, worth doing

I say 'in some respect' worth doing, because the fact that there is something that counts in favour of doing an action does not mean that it is, *all things considered* worth doing. For instance, although something counts in favour of having a doughnut every day (they taste great!), so that it is, in some respect, worth doing, it is probably not *all things considered* worth doing (the calories!). In such cases we can say that the action is, in some respect, worth doing, but not *all things considered* worth doing.

[14] To the extent that 'in virtue of' relation is an explanatory relation, Raz (perhaps against his own judgment) is so categorisable, cf. 'reasons are facts *in virtue of* which those actions are good in some respect.' (Raz 1999a, 22 emphasis added)
[15] As I did in § (I)4.2.

When is an act, in some respect, worth doing? It might be when it results in (or just is) something that is, in some respect, good, right, valuable, or merely desired.[16] For the purposes of this discussion we don't need to decide which of those it is – all we need note is that an act is, in some respect, worth doing if, *as a matter of fact*, it would result in something that is, in some respect, good, right, valuable, desired or what have you.

An important point to take from the last remark is this: whether or not an action is worth doing (in some respect, or *all things considered*) for some agent is an objective matter of fact that is, in particular, independent of their perspective. So an action is still worth doing if an agent is ignorant of the fact that it's worth doing. For instance: it's worth taking a different route home if there's traffic on one's usual route – and that's worth doing even if one thinks one's usual route is all clear (that it's worth doing doesn't mean that one will do it, it just means that if one did it, some 'good'[17] would come of it).[18] And an action isn't worth doing just because you think it is: you might throw away some milk because you assume, falsely, that it's gone off – throwing away the milk is not worth doing, although you may think that it is. So it's the way the world is that matters to whether or not some action is worth doing for some agent,[19] not the way she takes it to be.

## 3.3 Favourism about reasons to act

So far I've said that, according to this view, there is a reason to act if and only if an action is, in some respect, worth doing, and I've explained what it is for an action to be worth doing. All those authors to whom I attribute favourism about reasons to act would, I think, agree with those remarks.

Where my characterisation of favourism about reasons to act strays into the contentious is that it goes beyond these remarks to claim that reasons are actually *what makes* the action worth doing. This may be unpalatable to those who take the 'counts in favour of' relation to be primitive. For those that do balk at it I am happy for them to shrug it off (along with my

[16] Alvarez gives what she describes as a 'rough' characterisation, which, to my mind, is a neat representation of many of the different views. She notes that for there to be a fact that favours an agent doing some action that 'requires their having some motivation that would be served by acting in the way favoured… The motivation may be such things as desires, plans, long-standing projects or values. And it may be something the agent actually has, or something she would have if she reasoned properly from her current motivations.' (Alvarez 2016a, 10) This is what I mean by saying that an act is 'favoured' if it is, in some respect, worth doing.

[17] However we choose to understand 'good'…

[18] Cf. In a game of chess, a good move is a good move regardless of whether or not you've spotted it.

[19] Note that according to Broome-type conceptions of the 'counting in favour of' relation, acts are worth doing *because* of the way the world is.

categorisation of their theories), although that means leaving unanalysed a relation that some interpret differently.[20]

What really matters is that all of the theorists to whom I attribute favourism about reasons to act are committed to the following position, which, as I will show, is what really creates trouble for them:

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

This is clearly logically entailed by favourism about reasons to act. Moreover, I think all those to whom I attribute favourism about reasons to act who might balk at my characterisation of it (such as Parfit and Scanlon) would nonetheless agree with (FAV).

To be clear: (FAV) is weaker than favourism about reasons to act, in the sense that the latter entails the former, but the former does not entail the latter. My point is just that that (FAV) is enough to create trouble for favourism about reasons to act.

# 4 The problems for favourism

To see what the problems for (FAV) are, let's return to Sally and her non-existent bear. Since there isn't actually a bear chasing her, *ex hypothesi*, running isn't worth doing for Sally. There is nothing to be gained (for anyone) from Sally's running. Indeed, it's possible that she might fall and hurt herself, or attract the attention of actual bears by running. So *not* running is worth doing, but running is really *to no extent* worth doing.[21] Thus:

(F10) Sally's running was not at all worth doing.

This observation, combined with the *prima facie* reasonable claims set out above, results in three distinct problems for the favourist view: The Rational Action Problem; The Deliberate Action Problem; and The Psychological Reason Problem (for Favourism).

## 4.1 The Rational Action Problem

The Rational Action Problem is this: it is rational for Sally to run, so she has a reason to run, so there is a reason for her to run, so her running is, in some respect, worth doing, but her running is in no respect worth doing! Explicitly, the following claims are mutually inconsistent:

(F1) It was rational for Sally to run.

[20] See § (I) fn. 26 for different interpretations of the 'counting in favour of' relation.

[21] You could make a story in which running was, in some respect, worth doing, if you liked (Sally needs to lose a few pounds, say), but that's not my story. In my story Sally stands to gain nothing, and potentially lose much, from running (and, likewise, nothing of worth accrues to anyone else if Sally runs).

(F2) If it is rational for *A* to *φ* then some *p* was a reason *A* had to *φ*.

(F8) For any *p,* if *p* was a reason *A* had to *φ* then *p* was a reason for *A* to *φ*.

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(F10) Sally's running was not at all worth doing.

## 4.2 The Deliberate Action Problem

The Deliberate Action Problem is this: Sally ran deliberately, so she ran for a reason, so there was a reason for her to run, so her running is, in some respect, worth doing, but her running is in no respect worth doing![22] Explicitly, the following claims are mutually inconsistent:

(F4) Sally ran deliberately.

(F3) If *A φs* deliberately then *A φs* for a reason.

(F7) If *A φs* for a reason then there was a reason, *p*, for *A* to *φ*.

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(F10) Sally's running was not at all worth doing.

## 4.3 The Psychological Reason Problem (for Favourism)

The Psychological Reason Problem (for Favourism) is this: Sally's reason for running was, *inter alia*, that she believed that a bear was chasing her, so that was a reason for her to run, so her running was, in some respect, worth doing, but her running was to no extent worth doing![23] Explicitly, the following claims are mutually inconsistent:

(F5) Sally's reason for running was, *inter alia*, that she believed a bear was chasing her.[24]

(F9) For any *p,* if *A φs* for the reason that *p* then *p* was a reason for *A* to *φ*.

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

[22] A variant of this problem is Broome's (2013, 71) 'quick objection' to the thesis that acting rationally is 'responding correctly to reasons.'

[23] It's perhaps worth noting that this is an argument used by those who already accept favourism about reasons to act to show that Sally's reason for running can't be that she believed that a bear was chasing her (as set out in § (I)4.2). My point is that when we start our investigation of what reasons are from *prima facie* reasonable claims about them, without assuming favourism about reasons to act, the argument runs in the other direction – against favourism.

[24] Recall that I assume that: for any *p*, *A*'s reason for *φ*ing was that *p* if and only if *A φs* for the reason that *p* (see § (I)1.5) .

(F10) Sally's running was not at all worth doing.

## 4.4 The Experiential Reason Problem (for Favourism)

The Experiential Reason Problem (for Favourism) is this: Sally's reason for running was, *inter alia*, that she heard a bear-like sound, so that was a reason for her to run, so her running was, in some respect, worth doing, but her running was to no extent worth doing! Explicitly, the following claims are mutually inconsistent:

(F6) Sally's reason for running was, *inter alia*, that she heard a bear-like sound.

(F9) For any *p,* if *A φs* for the reason that *p* then *p* was a reason for *A* to *φ*.

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(F10) Sally's running was not at all worth doing.

# 5 Responses to the problems for favourism

How can a proponent of favourism about reasons to act respond to these problems? The options are limited. I don't think one could tolerably reject either that it's rational for Sally to run, or that Sally runs deliberately – so (F1) and (F4) are off the menu of potential responses.

It might be tempting to think we could reject (F10) and kill three problems with one rejection. We should resist this urge – to do so is ultimately just to change the notion of 'being worth doing' in a way that amounts to no more than a rejection of favourism about reasons to act by another name.[25]

So, (F1), (F4) and (F10) are off the menu. This brings me to the conclusion of my argument: if you want to preserve (FAV) then you have to reject at least one of the *prima facie* reasonable

[25] Maybe you want to say that Sally's belief that a bear is chasing her is what favours her action (Veli Mitova (2015, 2016) gives this a good go). Well, we could see how it *could* make running worth doing. If, for instance, Sally has a particularly odd constitution such that if she believes that a bear is chasing her and she doesn't run then she will have a heart attack or suffer some other unpleasantness. Then, in that case, her belief that a bear is chasing her would make running (in some respect) worth doing. Why? Because, given that she believes that a bear is chasing her, if she doesn't run she'll have a heart attack. But this isn't the situation that Sally is in. Sally is like you or I – she hears what sounds like a bear running, knows safety is nearby, so she runs. Perhaps you want to say that something does favour her running: the fact that there might be a bear nearby. But that fact doesn't make running worth doing – since *there isn't a bear nearby*. The fact that in close possible worlds a bear is chasing her does not make it worth running in this *actual* world. As it stands Sally has nothing to gain and quite a bit to lose from running. Running is not worth doing for her, although it is the most rational thing to do – and given the concept of favouring that we are working with, this just means that nothing favours her action.

claims from each of the following groups (and possibly two from group 3 – since one of the disjuncts is a conjunction), and accept the counterintuitive consequences of doing so[26]:

**Group 1**
- (F2) If it is rational for *A* to *φ* then some *p* was a reason *A* had to *φ*.
- (F8) For any *p,* if *p* was a reason *A* had to *φ* then *p* was a reason for *A* to *φ*.

**Group 2**
- (F3) If *A φs* deliberately then *A φs* for a reason.
- (F7) If *A φs* for a reason then there was a reason, *p*, for *A* to *φ*.

**Group 3**

EITHER:

- (F5) Sally's reason for running was, *inter alia*, that she believed a bear was chasing her. And;
- (F6) Sally's reason for running was, *inter alia*, that she heard a bear-like sound.

OR:

- (F9) For any *p,* if *A φs* for the reason that *p* then *p* was a reason for *A* to *φ*.[27]

For reference, here is a very much non-exhaustive account of which of these *prima facie* reasonable claims some different proponents of favourism about reasons to act reject (where: × = rejects):

| | Group 1 | | Group 2 | | Group 3 | |
|---|---|---|---|---|---|---|
| | (F2) | (F8) | (F3) | (F7) | (F5)&(F6) | (F9) |
| The Received View | × | | | × | | × |
| Stout (2009), Alvarez (2010), Parfit (2011), Littlejohn (2012) | × | | × | | × | |
| Hornsby (2008)[28] | | × | | × | × | |
| Schroeder (2008), Comesaña & McGrath (2014) | | × | | × | × | × |
| Dancy (2000, 2014), Davis (2005), Sandis (2009) | × | | | × | × | × |

**Table II-1:** How proponents of favourism respond to some problems for it

[26] Groups 1, 2 and 3 consist of the remaining premises of The Rational Action Problem, The Deliberate Action Problem, and The Psychological Reason Problem (for Favourism) & The Experiential Reason Problem (for Favourism), respectively.

[27] It's worth noting that (F9) entails (F7) – so rejecting the former serves as a response to both The Deliberate Action Problem and The Wrong Reasons Problem.

[28] I think that Hornsby's (2008) view with respect to the truth of (F9) has changed. In an earlier work she states that '*p* may be the agent's reason [for acting] even when it is false that *p*.' (2007, 299) If *p* were false it could not have been a reason for anyone to act because falsehoods don't count in favour of anything (that is not to say that *negations* don't favour anything – negations are facts that favour some things, but falsehoods aren't facts at all). Hornsby's earlier view is thus more closely aligned to the views of Schroeder (2008) and Comesaña & McGrath (2014). However in her later work she suggests that 'a condition of *φ*ing for the reason that *p*, when one believes that *p*, is that one knows that *p*.' (2008, 251) I think her idea in this later work is that one may act for a reason though there may be no reason for which one acts – this is something that, in § (I)1.5, I assumed could not happen. To the extent that that is her view, its nuances are not captured in my categorisation schema.

Of course, many of the proponents of favourism about reasons to act have good arguments as to why it is acceptable to reject these *prima facie* reasonable claims. For instance, Alvarez (2010, 142) defends her rejection of (F3) by noting that when an agent acts on a false belief we might well say that they act for no reason – someone who is running to catch a train might say, on discovering that it has been cancelled, 'You mean I ran all this way for no reason?'[29] Meanwhile Dancy (2000) suggests that we can explain away expressions like 'Sally's reason for running was that she believed that a bear was chasing her' as actually meaning 'Sally's reason for running was that a bear was chasing her, as she believed' – where the 'as she believed' is meant to be understood appositionally, in a manner that only qualifies what is said, rather than changing the meaning of it.

I am not seeking to refute these arguments here. My point is not that these arguments are wrong or that, more generally, there is no way to make the rejection of these *prima facie* reasonable claims intelligible My point is that rejecting these *prima facie* reasonable claims is a stance that demands some explanation – because, in doing so you are rejecting something that *on the face of it* seems reasonable. So that at least counts in favour of looking for a theory that doesn't commit one to such rejections.

# 6 Conclusion

The purpose of this chapter was to show the costs involved in accepting favourism about reasons to act. I have argued that if one wants to accept that view, one must reject several *prima facie* reasonable claims. This is a price that some are happy to pay; I would rather not.

[29] Alvarez put this particular example to me in a discussion of this point. By way of further response, Parfit (2001) and Alvarez (2010) would both suggest that we should say that both Sally and the person whose train was cancelled run for an *apparent* reason but not a *genuine* reason.

# (III)

# Acting for psychological reasons

*In which I show what it costs to think that the reason for which an agent acts is always a feature of their psychology. I show how 'psychologism about the reasons for which we act' clashes with some prima facie reasonable claims. In particular, I show that is inconsistent with the idea that we are often able to act for reasons that make our actions morally worthy and, more generally, worth doing.*

My friend has won a much-coveted award; I read about it in a newspaper so I call her up and congratulate her. It seems natural to say that my reason for congratulating her was that she had won an award. It also, I think, seems natural to say that my reason for congratulating her was that I read that she had won an award. Another example: if Jonathan sees someone who is alone and in trouble, he could, under the right circumstances, help her for the reason that she was alone and in trouble.[1]

These fairly anodyne observations create a lot of difficulty for *psychologism about the reasons for which we act*, which says that an agent's reason for acting can only ever be a feature of her psychology, and is probably the *de facto* account of what an agent's reason for acting is.

The purpose of this chapter is to set out a number of *prima facie* reasonable claims that psychologism about the reasons for which we act must reject.[2] This is not meant to be a conclusive argument against that view – it is only meant to show that accepting this view comes at some cost.

## 1 Some *prima facie* reasonable claims

### 1.1 My reason for congratulating my friend

I've already stated my first *prima facie* reasonable claim, but it's worth re-iterating; when I read in a newspaper of record that my friend had won an award and I consequently call her up to congratulate her, it is natural to say that:

(P1) My reason for congratulating my friend was, *inter alia*, that she had won an award.

It is, I think, similarly natural to say that:

[1] This example is from Dancy (2000).
[2] As the former chapter did for favourism about reasons to act.

(P2) My reason for congratulating my friend was, *inter alia*, that I read that she had won an award.

We should add to this observation two seemingly obvious remarks, whose importance will become clear (if it is not already) when we consider the problems for psychologism about the reasons for which we act:

(P3) The fact that my friend had won an award is not a feature of my psychology.

(P4) The fact I read that my friend had won an award is not a feature of my psychology.

## 1.2 Morally worthy actions

Suppose that Jonathan is on his way to the office when he encounters someone who is alone and in trouble. Taking all morally relevant features of the situation into account, without one thought too many, without any undue considerations of furthering his own ends or other less upstanding concerns, he duly comes to her aid. I submit to you that what he does is a morally worthy act:

(P5) Jonathan's act of helping the woman is morally worthy.

What does the moral worth of his action, and acts in general, come down to? One seemingly relevant consideration is this: if Jonathan had helped her only because he knew that she was very rich and would reward him amply for doing so, the moral worthiness of his act dissipates. As Julia Markovits notes, 'when we do the right thing because it happens to suit us, or happens to be in our interest, our action has no moral worth. This is intuitive. Morally worthy actions must be performed for the right reasons.'[3] (2010, 203)

So, we can do the right thing without doing something morally worthy if we don't do it for the right reasons. And what are the right reasons? They are the ones that make the action right:

> My action is morally worthy if and only if… I perform the action I morally ought to perform, for the reasons[4] why it morally ought to be performed. (Markovits 2010, 205)

> When I do the morally right thing, because it is the right thing to do, my reasons are the facts that make that action right. (Garrard and McNaughton 1998, 53)

I take this to be an intuitive view, and it provides us with the basis for our next *prima facie* reasonable claim:

[3] Markovits' original says 'the right (motivating) reasons' – I omit the parenthetical terminological remark only because I am trying to avoid this terminology (see § (I)4). On her account 'motivating reasons' are 'the reasons for which an agent acts' and 'normative reasons' are 'the reasons for an agent to act'.

[4] Here Markovits' original reads 'the (normative) reasons why' – I omit it the parenthetical terminology, again, to avoid terminological confusion (see fn. 3).

(P6) *A*'s *φ*ing is morally worthy only if *A φs* for a reason that makes *φ*ing right.[5]

So, what Jonathan does is morally worthy only given that he does it for a reason that makes it right.

## 1.3 What makes it right?

So what makes it right for Jonathan to help her? Why is it that helping her is the right thing to do? As Dancy notes, 'it is because she is in trouble that I ought to help her, not because I think she is in trouble.' (2000, 52)

Dancy's point is that it is the objective features of the world that determine whether or not an act is the morally right thing to do, and not features of the agent's psychology. It is right for me to recycle because it will help the environment – and that is so even if I don't know that it will help the environment. Likewise, it is right for Jonathan to come to the woman's aid because she is in trouble, and that would be so even if he had no idea that she was in trouble.[6]

Of course, sometimes features of our psychology might enter into the fray as moral considerations: you might say that regardless of whether or not she is in trouble, given that Jonathan thinks she is, the right thing for him to do is to try and help her. That is, you might say that not helping someone who you believe to be in trouble is wrong, regardless of whether or not they are in trouble. And in that case it is a feature of Jonathan's psychology that makes his action right. On this account his action is doubly right: both because helping her will save her from trouble and because, if he does so, he won't be neglecting someone he believes to be in trouble.

However, it's also quite possible that Jonathan's act is made right *only* by the objective features of the situation. Even if you think that *just* believing that she is in trouble could make helping her the right thing to do (regardless of whether or not she's in trouble), it's possible that given the rest of what he believes, believing that she is in trouble doesn't make it right for him to try to help her, although the fact that she is in trouble continues to do so. For instance, suppose that while he believes that she is in trouble he also (falsely) believes that if he were to try to help her it would only worsen her situation – in that case his belief that she is in trouble

[5] This is entailed by Markovits' bi-conditional. I use the weaker claim because it is sufficient for my purposes.

[6] I'm not doing an analysis of what it is for an action to be 'right' and certainly not of what 'ought' means. Even if there is an ambiguity between objective and subjective 'ought' claims, I think there is a clear and commonplace sense of 'right' actions that is independent of the agent's perspective – and that's the sense I am working with – and it is, importantly, in this sense of being 'right' that I think (P6) is an intuitive claim. If it weren't then it would, for instance, be unclear how one could do the right thing for the wrong reasons, or the wrong thing for the right reasons.

does not make it right for him to try to help her (because he thinks that by helping her it would make her worse off). However, since he could, in fact, help her without worsening her lot, helping her is still, in some obvious sense, the right thing to do.

While I don't suppose that Jonathan is in the situation so described (in particular, he does think that he could improve her lot by helping her), since I want to remove complicating factors, I'm going to suppose that Jonathan is in a situation such that it is only objective features of the circumstance that make his action right. I take it as a given that that is at least possible even if it is also possible that features of his psychology could make his action right.

To what end, all this convoluted reasoning? It is to make several points: objective (that is, worldly, not psychological) features of a situation can make an action right; and whilst features of an agent's psychology may also be able to make an action right, they don't need to – the reasons why it is right for some agent to do some action could have nothing to do with their psychology, and I'm saying that that is actually the case in the example of Jonathan and the troubled woman, as I construct it. Thus, we can, *ex hypothesi*, make the following *prima facie* reasonable claim:

(P7) No features of Jonathan's psychology make helping the woman right.

## 1.4 Acting for reasons that make it worth acting

If my friend wins an award and I congratulate her it will let her know that I've thought of her and am pleased for her – and she'll get some joy from that (and other things besides). Maybe I'll also get some joy from it too. In the sense discussed in § (II)3.2, congratulating my friend is something that is, for me, worth doing.

What makes congratulating my friend an act that is, for me, worth doing? It seems that it is things like this: the fact that she won an award, the fact that it would please her to be congratulated (given that she had the won the award), and so on. That is, facts about the way the world is make my act worth doing (including facts about my friend's psychology). Even if I didn't know these things, if I had no idea about her having won the award, congratulating her would still be worth doing – I just wouldn't do it. So let's say this:

(P8) The fact that my friend won an award makes congratulating her worth doing.

Now, here's a relevant *prima facie* reasonable claim about reasons that I want to put forward:

(P9) For any *p,* if *p* makes *A*'s *φ*ing worth doing then *A* could *φ* for the reason that *p*.[7]

A note on 'could': what is the modal concept we are working with here? I think there are intuitive grounds for a quite restrictive one, however, all that I need for the purpose of this argument is logical possibility: that is, if *p* makes *A*'s *φ*ing worth doing then it is *logically possible* that *A* could *φ* for the reason that *p*.

This seems, to me, like a claim that is hard to deny, but if you need some persuasion consider this: if you aren't doing something for reasons that make it worth doing, what reasons are you doing it for? The wrong ones? It just seems odd to me to think that there could be an action that is worth doing but which one could never do for reasons that make it worth doing. It is, seemingly, only when you do something for reasons that actually make it worth doing that you, 'do the right action for the right reason.' (Lord 2008, 2) Indeed, the very idea of acting 'for the right reasons' seems to depend upon the logical possibility (and probably something stronger than that) of acting for reasons that make one's action worth doing. The falsity of (P9) would imply that there could be some actions that are worth doing that one could never do 'for the right reasons'. I would suggest that that is, at least, a *prima facie* implausible view, so that (P9) is, at least, *prima facie* reasonable.

## 2 Psychologism about the reasons for which we act

Recall the following claim about reasons from Table I-4:

> *Psychologism about the reasons for which we act*: For any *p, p* is a reason for which *A φ*'d if and only if *p* is a feature of *A's* psychology that rationalises *φ*ing and explains (in the right way) why *A φ*'d.

Davidson's 'Actions, Reasons and Causes' (2001a) is probably the progenitor of this view. In addition, I think that this claim (or something sufficiently similar to it) is also advocated by Turri (2009), Gibbons (2010) and Mitova (2015, 2016) (although they don't articulate it in these terms). This claim is also perhaps most commonly associated with what I have called the 'Received View' (of which the former are *not* advocates (see Table I-6)).

[7] This is closely related to Bernard Williams' claim that: 'If there are reasons for action, it must be that people sometimes act for those reasons, and if they do, their reasons must figure in some correct explanation of their action.' (Williams 1981, 102) (See also Dancy 2000, 101; Smith 2004, 175; Hornsby 2007, 301; Raz 2009, 194; Hieronymi 2011, 415; Way and Whiting 2016, 214). I use this version because Williams (and others who express this claim) take favourism about reasons to act (see previous chapter) for granted – but someone who rejects that view can satisfy Williams's claim without difficulty. (P9) entails Williams' claim (given favourism about reasons to act), while providing a claim that one who rejects the latter will still also have to reject in order to endorse psychologism about the reasons for which we act (see § 3.2).

Why would anyone hold this view? I think that most philosophers who endorse it do so because they, either implicitly or explicitly, think that an agent's reason for acting must make their action rational and that only features of an agent's psychology can make their actions rational. However, the focus of this discussion is on the problems that arise from psychologism about the reasons for which we act, so I leave aside the reasons why one might endorse it for a later discussion.[8]

It is possible that some of those to whom I have attributed psychologism about the reasons for which we act (see Table I-6) might balk at my exact wording of it.[9] Nonetheless, I think all of those theorists share a commitment to the following, which is entailed by psychologism about the reasons for which we act,[10] and which is what really creates trouble for it:

(PSY) For any *p,* if *A φ*s for the reason that *p* then *p* is a feature of *A*'s psychology.[11]

# 3 The problems for psychologism

I present three problems for (PSY): The Moral Worthiness Problem; The Right Reasons Problem; and The Non-Psychological Reason Problem.

## 3.1 The Moral Worthiness Problem

The Moral Worthiness Problem is this: Jonathan's act of helping the woman was morally worthy, so he did it for reasons that made it right, so features of his psychology made it right, but no features of his psychology made it right! Explicitly, the following claims are mutually inconsistent:

(P5) Jonathan's act of helping the woman is morally worthy.

(P6) *A*'s *φ*ing is morally worthy only if *A φs* for a reason that makes *φ*ing right.

(PSY) For any *p,* if *A φ*s for the reason that *p* then *p* is a feature of *A*'s psychology.

(P7) No features of Jonathan's psychology make helping the woman right.

[8] In § (VIII), I will discuss, at length, the motivating argument for psychologism.

[9] In the same way that, say, Scanlon and Parfit might have balked at my characterisation of their views as 'favourism about reasons to act' (see § (II)3.3).

[10] For the sake of clarity: (PSY) is weaker than psychologism about the reasons for which we act, in the sense that the latter entails the former, but the former does not entail the latter. My point is that (PSY) is what causes psychologism about the reasons for which we act to face the problems it does.

[11] Cf. 'Psychologism… is the claim that the reasons for which we act are psychological states of ourselves.' (Dancy 2000, 98).

## 3.2 The Right Reasons Problem

The Right Reasons Problem is this: the fact that my friend has won an award makes congratulating her worth doing, so it's logically possible for me to congratulate her for the reason that she has won an award, so the fact that my friend has won an award must be a feature of my psychology, but it isn't! Explicitly, the following claims are mutually inconsistent:

(P8) The fact that my friend won an award makes congratulating her worth doing.

(P9) For any *p,* if *p* makes *A*'s *φ*ing worth doing then *A* could *φ* for the reason that *p*.

(PSY) For any *p,* if *A φ*s for the reason that *p* then *p* is a feature of *A*'s psychology.

(P3) The fact that my friend had won an award is not a feature of my psychology.

## 3.3 The Non-Psychological Reason Problem

The Non-Psychological Reason Problem is this: my reason for congratulating my friend was, *inter alia*, that she had won an award, so that must have been a feature of my psychology, but it isn't! Explicitly, the following claims are mutually inconsistent:

(P1) My reason for congratulating my friend was, *inter alia*, that she had won an award.

(PSY) For any *p,* if *A φ*s for the reason that *p* then *p* is a feature of *A*'s psychology.

(P3) The fact that my friend had won an award is not a feature of my psychology.

## 3.4 The Experiential Reason Problem (for Psychologism)

The Experiential Reason Problem (for Psychologism) is this: my reason for congratulating my friend was, *inter alia*, that I read that she had won an award in the newspaper, so that must have been a feature of my psychology, but it isn't! Explicitly, the following claims are mutually inconsistent:

(P2) My reason for congratulating my friend was, *inter alia*, that I read that she had won an award.

(PSY) For any *p,* if *A φ*s for the reason that *p* then *p* is a feature of *A*'s psychology.

(P4) The fact I read that my friend had won an award is not a feature of my psychology.

# 4 Responses to the problems for psychologism

What can a proponent of psychologism about the reasons for which we act do? It seems to me there is little to no choice in the matter. (P3) and (P4) are undeniable: of course facts about my friend or what I've read are not features of my psychology.

We constructed the Jonathan example so as to ensure the truth of (P7) (the claim that no features of Jonathan's psychology made his act right), so unless one wants to change the concept of 'making right', to a different one to that which I am using, one cannot reject (P7) (and if one were to change the concept, that would obviously just be a new (perhaps more solvable) problem, it would not be a solution to the problem I am posing). Similar remarks count against rejecting (P8).

What about (P5) (the claim that Jonathan's act is morally worthy)? Endorsing some form of moral anti-realism could allow one to claim that no acts are morally worthy, so that Jonathan's act isn't either, thereby rejecting (P5). This avoids having to reject (P6), but only by trivialising it – it is true only because the antecedent is never satisfied – so I think that a strategy that rejects (P5) is at least as *prima facie* implausible as just rejecting (P6).

So, leaving aside the possibility of rejecting (P5), there are no choices for the proponent of psychologism about the reasons for which we act; they must reject *all* of the following *prima facie* reasonable claims:

(P1) My reason for congratulating my friend was, *inter alia*, that she had won an award.[12]

(P2) My reason for congratulating my friend was, *inter alia*, that I read that she had won an award.[13]

(P6) *A*'s *φ*ing is morally worthy only if *A φs* for a reason that makes *φ*ing right.[14]

(P9) For any *p,* if *p* makes *A*'s *φ*ing worth doing then *A* could *φ* for the reason that *p*.[15]

Of course, one can perhaps put forward good arguments as to why these claims are nonetheless false, or why it is nonetheless acceptable to reject them. For instance, one might argue that (P9) is false, by providing cases in which one seemingly cannot do some action for a

[12] Rejecting (P1) solves The Non-Psychological Reason Problem.
[13] Rejecting (P2) solves The Experiential Reason Problem (for Psychologism).
[14] Rejecting (P6) solves The Moral Worthiness Problem.
[15] Rejecting (P9) solves The Right Reasons Problem.

reason that makes it worth doing.[16] Alternatively, perhaps one could construct counter-examples to (P6).[17]

My point, again, is not that one cannot argue against these claims, or that the arguments against these claims are wrong, rather it is just that these claims are *prima facie* reasonable so that rejecting them is a stance that demands some explanation – you are rejecting something that *on the face of it* seems reasonable. So, again, that at least counts in favour of looking for a theory that doesn't reject such *prima facie* reasonable claims.

# 5 Conclusion

As with the previous chapter's discussion of favourism about reasons to act, the purpose of this chapter was not to argue against psychologism about the reasons for which we act, but only to show the cost of accepting it, which is that one must reject (P1), (P6) and (P9), which are, as I have argued, all *prima facie* reasonable.

[16] Schroeder (2007, 33) gives the following example: The fact that there is a surprise party waiting at home for him makes going home early worth doing for Nate. However, Nate cannot go home early for the reason that there is a surprise party waiting (because it won't be a surprise party if he is aware of it, and he can't go home for the reason that there is a surprise party waiting for him if he isn't aware that there is).

[17] For instance, one might think that discovering the cure for cancer exclusively for the reason that one will be admired for having done so is nonetheless morally worthy. I would disagree, as, I think, many others would. Ultimately, I think, whether or not one rejects it depends upon whether or not one is a consequentialist about moral worthiness – noting that one can be a consequentialist about moral rightness without being a consequentialist about moral worthiness (cf. Mill: 'the motive has nothing to do with the morality of the action, though much with the worth of the agent.' (1863, 29)).

# (IV)

# Acting for what you believe

*In which I show what it costs to think that a reason for which an agent acts is the content of the belief they acted on. I show how 'deliberativism about the reasons for which we act' clashes with some prima facie reasonable claims about the factivity of reasons, the explanatory power of the reasons for which we act, the factivity of explanation and what an agent's reasons for acting are in Gettier cases. I set out which claims the proponent of this view must choose between rejecting.*

There is an account of the reasons for which an agent acts that aims to reconcile the idea that an agent who acts intentionally, deliberately and purposefully also acts for a reason with the idea that the reasons for which they act are often not features of their psychology. According to this account, which is what I have called 'deliberativism about the reasons for which we act', the considerations in light of which an agent acts, which is *what they believed* when they acted on their belief, are the reasons for which they act.

The problems for this view, as I shall show, arise from the *prima facie* reasonable claims that 'being a reason' is a seemingly factive property, that one's reasons for acting explain one's actions and that even when the considerations in light of which one acts are true, we are sometimes reluctant to call them the agent's reasons for acting.

## 1 Some *prima facie* reasonable claims about reasons

### 1.1 The factivity of reasons

If Sally is running because she mistakenly thinks that a bear is chasing her, we don't tend to say, 'her reason for running is that a bear is chasing her, even though one isn't'. One reason that we don't tend to say it is, I submit, that the expression sounds odd.[1] As Alvarez puts it, there is an 'air of paradox' about it:

> 'Othello kills Desdemona for the *reason* that Desdemona has been unfaithful to him, although she has not been unfaithful to him' sounds only marginally less paradoxical than 'Othello kills Desdemona because of the *fact* that Desdemona has been unfaithful to him, although she has not been unfaithful to him'. It is not that one cannot give meaningful interpretations to these expressions: we may hear them as conveying what the agent him- or herself would have said if asked about their reason... All the same, the fact that these expressions have an air of paradox

[1] Indeed, this claim was met with laughter when I put it to a seminar audience.

> and require this special interpretation are arguably explained by the thought that both 'the reason that' and 'the fact that' are factive operators. (Alvarez 2016b, 8)[2]

I think Alvarez's reading is correct: it is not that we can't find any meaning in the expression 'Sally's reason for running was that a bear was chasing her, even though one wasn't' – it's that the expression sounds odd, and we have to re-interpret it to make sense of it. Moreover, the fact that we *don't* say such things should tell us something about the applicability of the term to such cases.

I suggest, therefore, that the claim that '…is the agent's reason for acting' is a factive predicate is *prima facie* reasonable. Thus:

(D1) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* is the case.

## 1.2 The explanatory power of reasons

We can explain an agent's action by giving their reason for acting. Moreover, it is seemingly natural to think that when we so explain an agent's action, it is the agent's reason for acting *itself* that explains their action. When I say that Sally's reason for running was that she believed that a bear was chasing her it strongly seems to suggest (if not directly implies) that Sally ran *because* she believed that a bear was chasing her. Likewise, if I say that my reason for congratulating my friend was that she had won an award it suggests that I congratulated my friend *because* she had won an award.

The idea that an agent's reason for acting always explains their action is, as Lilian O'Brien notes 'very widely shared' (2015, 282). For instance:

> I act in light of those reasons. They are the reasons *why* I do what I do. (Dancy 2000, 103)

> When an agent acts for a (specific) reason that very reason is also the explanation (or at least part of the explanation) of why she did what she did. (Heuer 2004, 45)

> Taking something as one's reason, in acting on it, is taking it as an explanatory reason. (Setiya 2007, 36)

> A fairly standard way of linking reasons for action and explanations of action… is that *when someone acts for a reason then their reason for acting that way explains their acting that way*. (Stout 2009, 57 emphasis added)

> When there is a reason for which an agent acted then that reason explains (features in an explanation of) that action. (Raz 2011, 14)

[2] For others who take such remarks to *sound odd* see Unger (1978, 208), Scanlon (2014, 36), Dancy (2008a, 267) and Comesaña & McGrath (2014, 75) – the latter two are particularly worth noting since they advance a deliberative account and nonetheless recognize the oddness of such expressions.

I suggest that it is so widely held precisely because it matches our intuitions. The claim that something was one's reason for acting seems to be substitutable for the claim that one acted because of it. Thus, we have the second, *prima facie* reasonable claim for this discussion:

(D2) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* explains why *A φ*'d.

## 1.3 The factivity of explanation

Consider the following statements: I took my umbrella because it was raining, although it wasn't raining. The reason why the window broke was that a brick struck it, although a brick didn't strike it. That it's unfair explains why it's wrong, although it's not unfair. All of these statements sound strange to the point of unintelligibility. What makes them sound strange is that explanatory relations (whether causal or non-causal) are seemingly factive relations (and that is so whether they are picked out by 'because', 'explains' or 'reason why') – something cannot explain unless it is true.

I take this to be such an obviously *prima facie* reasonable claim that I won't defend its intuitiveness further:

(D3) For any *p*, if *p* explains anything then *p* is the case.

## 1.4 Gettier cases

Edmund Gettier (1963) introduced a now-familiar sort of character to epistemology: someone who has a justified, true belief that falls short of knowledge. Gettier characters are relevant to this discussion because, when an agent's justified belief is true by happy accident, what the agent believes (despite its truth) seemingly does not explain their action, as Hornsby demonstrates:

> Edmund…believes that the ice in the middle of the pond is dangerously thin, having been told so by a normally reliable friend, and…accordingly keeps to the edge. But Edmund's friend didn't want Edmund to skate in the middle of the pond (never mind why), so that he had told Edmund that the ice there was thin despite having no view about whether or not it actually was thin. *Edmund, then, did not keep to the edge because the ice in the middle was thin*. Suppose now that, as it happened, the ice in the middle of the pond was thin. This makes no difference. Edmund still didn't keep to the edge because the ice was thin. The fact that the ice was thin does not explain Edmund's acting, even though Edmund did believe that it was thin, and even though the fact that it was thin actually was a reason for him to stay at the edge. (Hornsby 2007, 251 emphasis added)

I think Hornsby's claim is intuitive: there is no sense in which Edmund's action was the consequence of the ice being thin, and, indeed, the ice could have been thick in the middle of the pond and Edmund would have stayed at the edge just the same. And given those observations, it seems hard to see how it could be true that Edmund stayed at the edge of the lake *because* the ice was thin. So, I suggest that the following claim is *prima facie* reasonable:

(D4) The fact that the ice was thin does not explain why Edmund stayed by the edge of the lake.

## 2 Deliberativism about the reasons for which we act

Recall the following claim about reasons from Table I-4:

> *Deliberativism about the reasons for which we act*: For any *p, p* is a reason for which *A* *φ*'d if and only if *p* is a consideration in light of which *A* *φ*s.

While deliberativism about the reasons for which we act is characterised by this bi-conditional, it may help the reader when we come to the problems with the account if we separate out the left-to-right and right-to-left readings (noting that the deliberative account is committed to both). Thus:

(DEL1) For any *p,* if *p* is a consideration in light of which *A* *φ*s then *p* is *A*'s reason for *φ*ing.

(DEL2) For any *p,* if *p* is *A*'s reason for *φ*ing then *p* is a consideration in light of which *A* *φ*s.[3]

It is easy enough to say that something is 'a consideration in light of which one acts', perhaps, but what does it mean?[4] Jonathan Dancy, who is probably the progenitor of contemporary deliberativism about the reasons for which we act, provides the following insight:

> The reasons for which we act are the considerations in the light of which we do what we do. These are the features which we take to tell in favour of so acting; they will figure prominently in our deliberation. (Dancy 2006, 123)

So, a condition on something's being a consideration in light of which one acts is that one should have taken it to 'favour' doing that action. Dancy, and other proponents of deliberativism, understand the favouring relation in the conventional way that I have already discussed (see § (I)4.4 and § (II)3.1) – as 'making, in some respect, worth doing'. So, a condition on something's being a consideration in light of which one acts is that one must take that thing to make one's action, in some respect, worth doing.

But this is not all there is to being a consideration in light of which one acts; one could take something to make an action, in some respect, worth doing and actually do that action without that thing being the consideration in light of which one did it. For instance, I could take the fact that it's pleasant outside to make going outside worth doing and I could actually go outside,

[3] It's worth noting that many who do not subscribe to deliberativism about the reasons for which we act nonetheless take (DEL2) to be true (e.g. Stout 2009; Alvarez 2010; Parfit 2011; Hyman 2015).

[4] These are what Scanlon (1998) calls 'operative reasons', what Olson and Svensson (2005) call 'deliberative reasons' and what Dancy (2000) and Schroeder (2008) call 'motivating reasons'.

but the consideration in light of which I do so might be, say, that I need to get lunch (I would still have gone outside if the weather had been dreadful).

What makes something a consideration in light of which one acts is not merely that one takes it to favour one's action, but that one *acts on it*. What that means may be a little opaque, but I think that we can say, by way of explication, that one acts on something that one took to make one's action worth doing if one acts *as a result* of taking it to make one's action worth doing.[5]

What does this all mean? Well, when I see rain outside, I believe that it's raining and I believe that my umbrella will keep me dry. I reason thus: it's raining, my umbrella will keep me dry, I want to stay dry (or judge it good or what have you), and so I had better take my umbrella. I take these considerations to favour taking my umbrella and I took my umbrella *as a result of* taking those considerations to favour doing so. It is these two conditions (taking some consideration to favour acting, and acting as a result of taking it to favour doing so) that make some consideration a consideration in light of which an agent acts. So much, then, for what it is to be a consideration in light of which one acts.

# 3 What Sally and Edmund took to favour acting

## 3.1 Sally's considerations

What were the considerations in light of which Sally ran? That is, what was it that she took to favour running (that is, to make running, in some respect, worth doing)? Well, Sally heard something that sounded like a bear, and her thought process, albeit a quick one, must have been something like: 'A bear's coming! The safe-house is nearby – I can make it. I had better run!' So, amongst the considerations that Sally took to favour running was *that a bear was chasing her*. Of course, a bear wasn't chasing her, but that doesn't prevent it from being a consideration in light of which she acted, any more than the fact that a bear wasn't chasing her prevented her from believing that one was.

In order for one to take something to favour some action, for it to be a consideration in light of which one acted, it is, as I have noted, enough that one believes it, believes that it favours one's action (in the sense of making it, in some respect, worth doing) and acts on those beliefs. In particular, it is not necessary that one know it, or even that it be the case, because the belief operator, unlike knowledge, is not factive.

[5] The natural interpretation of 'as a result' here is an explanatory one – but I want to avoid such a commitment since some (e.g. Dancy 2000) don't think the 'taking' is explanatory, but is merely an 'enabling condition' for the action.

This is a noteworthy upshot of the deliberative account: an agent can act in light of a consideration that was false. And this is what Sally does: Sally takes *that a bear was chasing her* to favour her action, and acts in the light of that consideration.

(D5) A consideration in light of which Sally ran was that a bear was chasing her.

Was the fact that she believed that a bear was chasing her a consideration in light of which Sally ran? It was not. Sally does not think to herself: 'Hmmm, I believe that a bear is chasing me. That I believe that a bear is chasing me makes running worth doing (regardless of whether or not a bear is chasing me…), so I had better do it.' It would be odd for Sally to reason this way, because it's only in unusual circumstances that we take considerations about *our own psychology* to favour our actions.

These remarks relate to remarks made in § (I)4.2, that are perhaps worth revisiting. Recall that Sam believes that the security services are trying to read her mind. She takes *that the security services are trying to read her mind* to favour wearing a foil hat (she takes it to be, in some respect, worth doing). However, knowing that she has a delusional disorder, she also takes *that she believes that the security services are trying to read her mind* to favour going to see a doctor. Sam deliberates about the way she takes the world to be as well as the way she takes her mind to be.

Sally is not like Sam – when Sally decides to run it's not because she has deliberated about her own psychology – what Sally takes to favour her action is the way world is, according to her. If Sally thought that *merely* believing that a bear was chasing her would make running worth doing *regardless of whether or not a bear was chasing her*, then it could be a consideration in light of which she runs. But, *ex hypothesi*, that isn't what Sally thinks, what Sally takes to favour running is *that a bear is chasing her* (even though no bear is actually chasing her) – Sally is just mistaken about what makes running worth doing, because nothing makes it worth doing.

The point is that we should not infer from the fact that Sally runs because she believes that a bear is chasing her that what she deliberates about is her mental states rather than their propositional contents. In short:

(D6) That she believed that a bear was chasing her was not a consideration in light of which Sally ran.

What about the fact that Sally heard a bear-like sound? Is that a consideration in light of which she ran? Maybe you want to say that Sally's reasoning goes like this 'That sounded like a bear!

There must be a bear coming! I had better run!' Her reasoning might go like that. But that doesn't make hearing a bear-like sound something she took to favour running.

Why not? Because she doesn't think that hearing a bear-like sound makes running worth doing. Sally doesn't think: 'if I hear a bear-like sound I should run, regardless of whether or not a bear is chasing me.' Hearing a bear-like sound is what makes her believe that a bear is chasing her, but it's *that a bear is chasing her* that she takes to favour running. That is, if hearing a bear-like sound plays *any* role in her deliberation it is in helping her decide what to believe; but once she has settled on believing that a bear is chasing her it is *what she believes* that she takes to favour her action, not what she took to favour believing it. So:

(D7) That she heard a bear-like sound was not a consideration in light of which Sally ran.

### 3.2 Edmund's considerations

What were the considerations in light of which Edmund stayed by the edge of the pond? That is, what was it that he took to favour staying by the edge of the pond and on the basis of which he did so?

Well, his thought process might have gone (if somewhat elaborately) like this: 'The ice is thin in the middle. If I skate there it might crack, I might fall through. That would be dreadful, perhaps fatal. I'd better just stay at the side.' Assuming that this (or something in its vicinity) is how his reasoning went, the things that Edmund took to favour his action, and which made the difference to what he did, were things like *that the ice was thin*, *that skating on thin ice is dangerous* and so on. Thus, we can say:

(D8) A consideration in light of which Edmund stayed by the edge was that the ice was thin.

## 4 The problems for deliberativism

The problems for deliberativism about the reasons for which we act will no doubt be obvious from the remarks of the previous sections. Nonetheless, it is worth being explicit about them. What follows are four distinct problems for the deliberative account: The False Reasons Problem; The Explanatory Reasons Problem; The Deliberative Gettier Problem; and The Psychological Reason Problem (for Deliberativism).

## 4.1 The False Reasons Problem

The False Reasons Problem is this: A consideration in light of which Sally ran was that a bear was chasing her, so it was her reason for running, so a bear was chasing her, but no bear was chasing her! Explicitly, the following claims are mutually inconsistent:

(D5) A consideration in light of which Sally ran was that a bear was chasing her.

(DEL1) For any *p,* if *p* is a consideration in light of which *A φ*s then *p* is *A*'s reason for *φ*ing.

(D1) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* is the case.

(D9) It is not the case that a bear was chasing Sally.

## 4.2 The Explanatory Reasons Problem

The Explanatory Reasons Problem is this: A consideration in light of which Sally ran was that a bear was chasing her, so it was her reason for running, so it explains why she ran, so a bear was chasing her, but no bear was chasing her! Explicitly, the following claims are mutually inconsistent:

(D5) A consideration in light of which Sally ran was that a bear was chasing her.

(DEL1) For any *p,* if *p* is a consideration in light of which *A φ*s then *p* is *A*'s reason for *φ*ing.

(D2) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* explains why *A φ*'d.

(D3) For any *p*, if *p* explains anything then *p* is the case.

(D9) It is not the case that a bear was chasing Sally.

## 4.3 The Deliberative Gettier Problem

The Deliberative Gettier Problem is this: A consideration in light of which Edmund stayed by the edge of the pond was that the ice was thin in the middle, so that was his reason for staying by the edge, so it explains why he stayed by the edge, but it doesn't explain why he stayed by the edge! Explicitly, the following claims are mutually inconsistent:

(D8) A consideration in light of which Edmund stayed by the edge was that the ice was thin.

(DEL1) For any *p,* if *p* is a consideration in light of which *A φ*s then *p* is *A*'s reason for *φ*ing.

(D2) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* explains why *A φ*'d.

(D4) The fact that the ice was thin does not explain why Edmund stayed by the edge of the lake.

## 4.4 The Psychological Reason Problem (for Deliberativism)

The Psychological Reason Problem (for Deliberativism) is this: Sally's reason for running was, *inter alia*, that she believed a bear was chasing her, so that was a consideration in light of which she ran, but it wasn't![6] Explicitly, the following claims are mutually inconsistent:

(F5) Sally's reason for running was, *inter alia*, that she believed a bear was chasing her.[7]

(DEL2) For any *p,* if *p* is *A*'s reason for *φ*ing then *p* is a consideration in light of which *A φ*s.

(D6) That she believed that a bear was chasing her was not a consideration in light of which Sally ran.

## 4.5 The Experiential Reason Problem (for Deliberativism)

The Experiential Reason Problem (for Deliberativism) is this: Sally's reason for running was, *inter alia*, that she head a bear-like sound, so that was a consideration in light of which she ran, but it wasn't! Explicitly, the following claims are mutually inconsistent:

(F6) Sally's reason for running was, *inter alia*, that she heard a bear-like sound.

(DEL2) For any *p,* if *p* is *A*'s reason for *φ*ing then *p* is a consideration in light of which *A φ*s.

(D7) That she heard a bear-like sound was not a consideration in light of which Sally ran.

# 5 Responses to the problems for deliberativism

What can the proponent of this account do? Rejecting (D9) is not an option: a bear wasn't chasing Sally – that much we know for sure. Rejecting any of (D5), (D6), (D7) or (D8) (the claims

[6] This argument is often (e.g. Dancy 2000, 124–25; Alvarez 2016b, 9) used by proponents of (DEL2) to argue against the idea that Sally's reason for running was that she believed that a bear was chasing her. My point is that it is *prima facie* reasonable to suggest that Sally's reason for running was that she believed that a bear was chasing her – so if one's theory of reasons commits you to rejecting it, it is at some cost that it does so.

[7] This is from § (II)1.3, where we first considered this example.

about the considerations in light of which Sally and Edmund did or did not act) is also not an option – the truth of (D5) to (D8) just follows from what it is to be a consideration in light of which one acts (and the construction of the examples). Of course one could change the notion of consideration that we are working with, but that would be to change the account, it wouldn't solve the problem for deliberativism about the reasons for which we act.

So, the proponent of deliberativism about the reasons for which we act must reject:

(F5) Sally's reason for running was, *inter alia*, that she believed a bear was chasing her. [8]

(F6) Sally's reason for running was, *inter alia*, that she heard a bear-like sound.[9]

(D1) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* is the case.[10]

And, EITHER:

(D3) For any *p*, if *p* explains anything then *p* is the case.[11] And;

(D4) The fact that the ice was thin does not explain why Edmund stayed by the edge of the lake.[12]

OR:

(D2) For any *p*, if *p* was *A*'s reason for *φ*ing then *p* explains why *A φ*'d. [13]

As ever, one can, perhaps, construct compelling arguments as to why these claims are, in spite of their *prima facie* reasonableness, nonetheless false. For instance, Dancy (2000) suggested that rejecting (D3), and insisting that there is such a thing as non-factive explanation, might be the appropriate response to The Explanatory Reasons Problem. This view has proved to be less than compelling; indeed, the claim that explanation could be non-factive is apparently so unpalatable a thesis that even Dancy has now abandoned it, yielding to what he describes as 'a barrage of criticism'.[14]

The rejection of (D2) is now the more favoured approach amongst deliberativists.[15] The main strategy for doing so appears to be this: given that an agent's reason for acting is an intentional object (*qua* what the agent believes), it is not the sort of thing that can do

[8] Rejecting (F5) solves The Psychological Reason Problem (for Deliberativism).
[9] Rejecting (F6) solves The Experiential Reason Problem (for Deliberativism).
[10] Rejecting (D1) solves The False Reasons Problem.
[11] Rejecting (D3) solves The Explanatory Reasons Problem.
[12] Rejecting (D4) solves The Deliberative Gettier Problem.
[13] Rejecting (D2) solves both The Explanatory Reasons Problem and The Deliberative Gettier Problem.
[14] See Dancy (2014). Although, Comesaña & McGrath (2014) appear to have picked up the non-factive baton.
[15] (E.g. Stoutland 1998; Davis 2003, 2005; Sandis 2013; Dancy 2014)

explaining, so (D2) can't be true. The problem, it seems to me, is that this line of argument reasons to its conclusion by taking (DEL1) as given, when that is precisely what is in contention.[16]

Nonetheless, as with the previous discussions, what is at issue here is not whether or not an argument can be given that makes rejecting (D2) (and (F5), (F6) and (D1)) tolerable, my point is just that it counts against deliberativism about the reasons for which we act that such an argument needs to be given in the first place. If your theory has some counter-intuitive consequences its counter-intuitiveness doesn't make it wrong, but it certainly counts against it when compared to a more intuitive theory.

# 6 Conclusion

This chapter has shown that the deliberative account must reject a number of *prima facie* reasonable claims, as previous chapters did for favourism about reasons to act and psychologism about the reasons for which we act. The next chapter considers whether or not the difficulties faced by these three accounts should persuade us that there are just several, irreconcilable concepts of reason at play.

[16] That is, it concludes that an agent's reason for acting can't explain their action by assuming it is an intentional object, but this begs the question, which is, in part, about whether or not the agent's reason for acting *is* an intentional object.

# (V)

# On the plurality of reasons

*In which I explain what a pluralist theory of reasons is and why 'going plural' is not a panacea. I suggest that a given reason expression could have more than one sense, and I show how we can accommodate theories of reasons that accept that idea, i.e. pluralist theories of reasons, in our categorisation schema. I discuss some examples of pluralist theories from the literature. I show how pluralist theories can solve some of the problems discussed in the previous chapters. I explain why pluralism is not, however, enough, and I suggest that our investigation should go beyond favourism, psychologism and deliberativism.*

The previous three chapters set out some of the main problems for the most popular claims from each of the three families of claims about reasons. In doing so they painted a bleak picture of contemporary theories of reasons; as Dancy notes (paraphrasing Aristotle) the theories 'leave one saying things that nobody would say unless defending a theory.' (2008a, 267) So what is to be done? One possible solution, which I wish to consider now only so that we may set it aside, is to think that there are different *senses* of a given reason expression.

Up until now all of the theories I have considered have been univocal; that is, they have (as I noted in § (I)1) all assumed that a given reason expression always picks out reasons of the same kind. Homonyms, I said, are the exception and not the rule. But what if reason expressions are homonyms of some sort, picking out kinds of things that are confusingly similar, but nonetheless distinct? Perhaps, that is, there is a *sense* in which Sally runs for a reason and a *sense* in which she doesn't run for a reason (or runs for no reason). Similarly, perhaps if there is no milk at home but I believe that there is, then there is a *sense* in which there is no reason for me to buy milk and a *sense* in which there is a reason for me to buy milk. Perhaps, the same reason expression can have different *senses*; perhaps, that is, the same reason expression can be used to pick out different kinds of reason.

A theory that admits that a single reason expression can have different senses is a *pluralist* theory of reasons. The purpose of this chapter is to explain what pluralist theories of reasons are, why one would adopt pluralism, and to show that even if we adopt a pluralist theory of reasons, we should look for a new account of the reason-relation, beyond favourism, deliberativism and psychologism.

# 1 The sense of an expression

For expositional purposes I want to stress the distinctions we are now working with: there are (i) different reason expressions, (ii) different kinds of reason they pick out and, now (iii) different *senses* of a reason expression.

In § (I)1, I discussed, at length, the possibility that different reason expressions might pick out different kinds of reason. However, I assumed that each reason expression picks out only one kind of reason. Now, I am allowing that the *same* reason expression may pick out *different* kinds of reason – and to the extent that it does, we can say that the reason expression has different *senses*. That is, a single reason expression has different *senses* if and only if it can be used to pick out different kinds of reason (and two kinds of reason are different if and only if the conditions for being a reason of each kind differ).

A further, important, clarification is to note that saying that a reason expression has different senses is not the same as saying that it picks out a 'disjunctive' kind of reason. If, for instance, 'a reason there is to act' just picks out one kind of reason, which happens to be disjunctive (in the sense that the conditions for being a reason of that kind are disjunctive[1]), then there could not be a sense in which something is and a sense in which it isn't a reason to act – there is just one sense of that reason expression, even though the conditions for its application are disjunctive. In contrast if a single reason expression can be used to pick out two different kinds of reason, there are two senses to that expression, and this is so even if both of the kinds of reason picked out are non-disjunctive.[2]

# 2 Expanding the categorisation schema

To account for the possibility of a plurality of senses for any given reason expression, we need to expand the categorisation schema set out in Table I-3 (and, indeed, expanding the schema in this way can be used to further explicate what we mean by there being different senses).

[1] I haven't considered any disjunctive conditions for being a reason (I am not convinced that anyone holds what I would call a disjunctive theory). However, we would have a disjunctive claim about reasons to act if, for instance, we said: '*p* is a reason for *A* to *φ* if and only if either *p* makes *A*'s *φ*ing, in some respect, worth doing or *A* takes *p* to make *A*'s *φ*ing, in some respect, worth doing.' This would be a disjunctive, but nonetheless univocal, account of the expression 'a reason there is to act' (it is disjunctive between the favourist and deliberativist conditions).

[2] In the event that this is not clear, consider the following: there is a sense of the word 'bat' according to which that which a hitter in a baseball game uses is a 'bat', and a sense in which it isn't a 'bat' (it's not a winged mammal). The expression 'a bat' has two senses, each picking out a non-disjunctive kind of thing. In contrast '(sporting) bat ∨ (animal) bat' is a disjunctive expression (you could say that it picks out a 'disjunctive kind of thing' if you believe in disjunctive kinds) that has only one sense – it's not the case that something could be both '(sporting) bat ∨ (animal) bat' and not '(sporting) bat ∨ (animal) bat' (that is, if something is either a sporting bat or an animal bat then is a '(sporting) bat ∨ (animal) bat'). My point is just that being disjunctive is unrelated to having several senses.

Table V-1 is a categorisation schema that allows for two senses of a given reason expression and thereby provides a way to represent both univocal and pluralist[3] theories of reasons.

| **Reason expression** | **Sense *A*** | **Sense *B*** |
|---|---|---|
| For any *p*, *p* is a reason for *A* to *φ*… | *Claim* | *Claim* |
| For any *p*, *p* is a reason for *A's φ*ing… | *Claim* | *Claim* |
| For any *p*, *p* is a reason *A* has to *φ*… | *Claim* | *Claim* |
| For any *p*, *p* is *A*'s reason for *φ*ing… | *Claim* | *Claim* |

**Table V-1:** A categorisation schema that accommodates pluralist theories of reasons

Recall that a reason expression only picks out different kinds of reason if the conditions between different *senses* of the expression differ (if they don't then the 'different' senses both pick out the same kinds of reason, in which case there is really only *one* sense). So, a univocal theory is represented in Table V-1 by providing the *same* conditions under both senses. For instance, the 'Received View', which is a univocal theory that is classified in the original schema as (F, F, F, P), would be as follows in this new schema:

| **Reason expression** | **Sense *A*** | **Sense *B*** |
|---|---|---|
| Reasons there are to act | Favourism | Favourism |
| Reasons for acting | Favourism | Favourism |
| Reasons one has to act | Favourism | Favourism |
| Reasons for which one acts | Psychologism | Psychologism |

**Table V-2:** The, univocal, 'Received View' represented in the new schema

There is only one sense to each expression in the 'Received View', hence the claims about each reason under each 'sense' in the categorisation schema are the same. We can enrich our 4-tuple descriptions to represent the possibility of theories with multiple senses of a given reason expression by introducing a '/' to denote alternate senses. So: (F, F, F, P) ≡ (F/F, F/F, F/F, P/P).

Now, in contrast to this univocal theory, a pluralist theory of reasons is any theory that, for *some* reason expression, makes a different claim under each sense. Here is an example pluralist theory represented using this schema:

[3] Or at least 'dual' sense theories – I suppose it's possible that an expression could have more than two senses, but I don't consider that here.

| Reason expression | Sense *A* | Sense *B* |
|---|---|---|
| Reasons there are to act | Favourism | Deliberativism |
| Reasons for acting | Favourism | Favourism |
| Reasons one has to act | Favourism | Favourism |
| Reasons for which one acts | Favourism | Favourism |

**Table V-3:** An example pluralist theory

The 4-tuple description of this theory is (F/D, F/F, F/F, F/F). This is not, to my knowledge, one that anyone advocates – I use it only to indicate what a pluralist theory looks like in this schema. This theory is pluralist with respect to reasons there are to act; it takes 'a reason there is to act' to pick out either a fact that makes the act, in some respect, worth doing, or something that the agent took to make the act, in some respect, worth doing. But it is univocal with respect to all the other reason expressions.

In what follows I want to briefly represent (what I take to be) the two main 'pluralist' theories of reasons that have been considered in the literature to date.

# 3 Favourist/Deliberativist (F/D) pluralism

## 3.1 Objective and subjective reasons

An increasingly common response to cases like Sally's mistake about the bear or my ignorance about my lack of milk is to distinguish between *objective* and *subjective* kinds of reason (e.g. Stoutland 2007; Schroeder 2007; Markovits 2011; Vogelstein 2012; Whiting 2014). An *objective* reason is something that (in my parlance) makes one's action, in some respect, worth doing, whereas a *subjective* reason is something that (again, in my parlance) the agent took to make their action, in some respect, worth doing.[4]

So, to give some examples, what Sally takes to make running worth doing (*that a bear is chasing her*) is a *subjective* reason for her to run but not an *objective* reason for her to run. And, in contrast, the fact that I am out of milk is an *objective* reason for me to buy more (because it makes it, in some respect, worth doing) but not a *subjective* reason (because I believe that I have plenty – I don't take anything to make buying milk, in some respect, worth doing). If, on the other hand, it is raining and I believe that it is raining then the fact that it is raining is both an *objective* reason and a *subjective* reason for me to take my umbrella

[4] An alternative vernacular for the subjective/objective reasons distinction is talk of first-person and third-person reasons (respectively).

(because it both makes taking an umbrella, in some respect, worth doing and I believe that it does).[5]

It should perhaps be clear that, given these definitions, an objective reason is the kind of reason that favourists take reason expressions to pick out (i.e. the sort of thing that makes actions, in some respect, worth doing), and a subjective reason is the kind of reason that deliberativists take reason expressions to pick out (i.e. the sort of thing that agents take to make their actions, in some respect, worth doing).

## 3.2 A pluralist theory

So far the distinction between objective and subjective reasons is merely terminological. One could make this distinction and yet retain a univocal theory of reasons if one were to say that no single reason expression can be used to pick out either objective or subjective reasons.[6] However, some explicitly invoke the distinction between objective and subjective reasons so as to offer a pluralist theory of reasons. For instance, Vogelstein says of Parfit's (2011) well-known 'snake' example[7]:

> There seems to be a sense in which there is a reason for you to run away (since you believe that running away will save your life), and a sense in which there is no reason for you to run away (since no good will come of it). That is, there is a subjective reason, but no objective reason, for you to run away. Likewise, there is a sense in which there is a reason for you to stand still (since it will save your life), and a sense in which there is no reason for you to stand still (since you believe nothing to suggest that any good will come of it). That is, there is an objective reason, but no subjective reason, for you to stand still. (Vogelstein 2012, 241)

Vogelstein is saying is that there are two senses to being *a reason there is to act*. One sense, he suggests, corresponds to objective reasons, the other corresponding to subjective reasons. So, since he thinks one of the reason expressions has two senses, he is offering what I have called a 'pluralist' theory of reasons. And according to his pluralist theory of reasons one sense of the expression, 'a reason there is to act' is favourist and the other sense is deliberativist.

---

[5] If one assumes that intentional objects are propositions and that true propositions are facts – otherwise the ontology of subjective and objective reasons is different, so one and the same thing cannot be both an objective and a subjective reason. That being so, in this example I would still have an objective and a subjective reason, they would just be different things on account of the ontological difference between these kinds of reason.

[6] For instance, I think that Schroeder (2008) invokes the objective and subjective reason distinction only to say that the expression 'a reason there is to act' picks out objective reasons, whereas the expression 'a reason one has to act' picks out subjective reasons (this is, I think, also Dancy's (2012) interpretation of Schroeder's view), so his theory is not pluralist, it just distinguishes between the kinds of reason picked out by these expressions (which are normally taken to be co-extensive (see § (I)1)). For the classification of Schroeder's view see Table I-6.

[7] Here, for reference, is Vogelstein's version of this example: 'While walking in a desert, you have angered a poisonous snake. You believe that running away will save your life, and believe nothing to suggest otherwise. As it turns out, however, you must stand still in order to save your life, as this snake will attack moving targets.' (Vogelstein 2012, 241)

Vogelstein is not alone. Although they don't talk in terms of 'objective' and 'subjective' reasons, both Hyman (2011) and Locke (2015) offer a pluralist account of the expression 'the agent's reason for acting,' of precisely this kind. They both hold (in my parlance) that there is a sense of 'the agent's reason for acting' that picks out a consideration in light of which the agent acted (that is, a *subjective* reason that they acted on), and a different sense that picks out a fact that makes their action, *all things considered*, worth doing, and explains it in the right way (that is, an *objective* reason that explains their action in the right way).[8] Thus, both Hyman and Locke think that one of the senses of being a reason for which an agent acts corresponds to *favourism about the reasons for which we act* and another sense corresponds to *deliberativism about the reasons for which we act*.

So there are pluralist theories that mix favourism and deliberativism. I don't want (or need) to categorise them all, so, for the sake of argument let's just characterise a pluralist theory of this kind as pluralist about every reason expression (i.e. (F/D, F/D, F/D, F/D)):[9]

| **Reason expression** | **Sense *A*** | **Sense *B*** |
|---|---|---|
| Reasons there are to act | Favourism | Deliberativism |
| Reasons for acting | Favourism | Deliberativism |
| Reasons one has to act | Favourism | Deliberativism |
| Reasons for which one acts | Favourism | Deliberativism |

**Table V-4:** Pure F/D pluralism

# 4 Favourist/Psychologist (F/P) pluralism

Michael Smith (1987, 1994) is typically taken to hold what I have called the 'Received View' (see Table V-2), which is a combination of, *inter alia*, favourism about reasons to act and psychologism about the reasons for which we act. I think that this is a misreading of Smith. In what follows I want, briefly, to make the case that Smith argues for a pluralist theory of reasons.

## 4.1 The misinterpretation of Michael Smith

Firstly, although Smith's definition of 'normative reasons' in these works is loose,[10] we can plausibly treat what he refers to as 'normative reasons' as what we have already called an

[8] Although it is perhaps worth noting that they differ in their views of what it takes for such a fact to explain an action *in the right way* (Hyman thinks it is knowledge, Locke thinks it is an explanatory chain).
[9] Some theories may not be pluralist about every expression, in the way that this one is – but that isn't very important here.
[10] (See Smith 1987, 39)

'objective reason' (that is, something that makes an action, in some respect, worth doing), which is, you will recall, what favourists take reasons to be. Smith's usage of the term 'normative reason' is thus consistent with its contemporary usage.[11]

In contrast, Smith's usage of the term 'motivating reason' is at odds with contemporary usage.[12] Smith argues that motivating reasons are all psychological states of the agent that rationalise their action (that is, motivating reasons are the sorts of thing that psychologism takes reasons to be), and this is because he *defines* motivating reasons such that:

> The distinctive feature of a motivating reason to *φ* is that in virtue of having such a reason an agent is in a state that is *potentially explanatory* of his *φ*ing. (Note the 'potentially'. An agent may therefore have a motivating reason to *φ* without that reason's being overriding.) (Smith 1987, 38 emphasis in original)

It is typical to interpret Smith's remarks about motivating reasons to be about 'the agent's reason for acting', and hence to take Smith to be arguing for a univocal psychologism about the agent's reason for acting[13] – this, I think, is a mistake. For one, the fact that he talks about motivating reasons being only *potentially* explanatory should already tell us that he isn't talking about the agent's reason for acting – the agent's reason for acting is something that is generally taken to be (and Smith certainly takes it to be) *actually* explanatory (see § (IV)1.2).

Furthermore, the fact that Smith talks about motivating reasons to *φ* should be indicative of the fact that the kind of reason he is talking about is such that reasons of that kind are (in the sense introduced in § (I)1) *independent* of the actions for which they are reasons. That is, the 'to' preposition makes clear that something could be a motivating reason *to* do something even if one does not do it. Now, since we know[14] that reasons of the kind picked out by the expression 'the agent's reason for acting', are *dependent* on the actions they are reasons for,[15] when Smith refers to 'motivating reasons to *φ*' he can't be talking about the agent's reason for acting.

Thus, contrary to the typical interpretation, when Smith talks about 'motivating reasons' he isn't necessarily talking about those reasons picked out by the expression 'the agent's reason for acting'. So, when he says that motivating reasons are features of the agent's psychology, he is likewise not necessarily advocating psychologism about the reasons for which we act. I think that in saying what he does, Smith isn't making a claim about any particular reason expression – he is just defining the term 'motivating reason'.

[11] See § (I)4.1 for a discussion of the contemporary usage of 'normative' and 'motivating' reasons.
[12] This a point that both Darwall (2003, 442–43) and Setiya (2007, 30) make.
[13] Which is, indeed, why he is associated with the Received View.
[14] See § (I)1.5.
[15] In the sense that its reason-hood depends on the occurrence of the action for which it is a reason.

The right way to understand Smith's usage of the terms 'normative' and 'motivating' reasons is, I suggest, as different senses of any given reason expression. For instance, I take such a pluralist conception to be the most natural way to construe the following remarks:

> The claim that *A* has a reason to *φ* is ambiguous. It may be a claim about a *motivating* reason that *A* has or a claim about a *normative* reason that A has. (Smith 1987, 38 emphasis in original)

Smith is saying that the expression 'a reason *A* has to *φ*' is ambiguous between two senses: being a normative reason (which is the favourist kind of reason, on Smith's definition) and being a motivating reason (which is a psychologistic kind of reason, on Smith's definition). That is, Smith is recommending a pluralist account of that expression, and therefore, a pluralist theory of reasons. Moreover, I think, Smith's pluralism extends to *all* the reason expressions.[16]

## 4.2 Another kind of pluralism

That having been said, Smith's actual theory is largely irrelevant to our primary concern. The remarks above are mainly intended as context (if somewhat polemical); all that we need to take from this discussion is that the materials for another pluralist theory of reasons are already out there – one which takes all reason expressions to have two senses, one favourist and the other psychologist (i.e. (F/P, F/P, F/P, F/P)). We can represent such a theory in our revised schema as follows:

| **Reason expression** | **Sense *A*** | **Sense *B*** |
|---|---|---|
| Reasons there are to act | Favourism | Psychologism |
| Reasons for acting | Favourism | Psychologism |
| Reasons one has to act | Favourism | Psychologism |
| Reasons for which one acts | Favourism | Psychologism |

**Table V-5:** Pure F/P pluralism

# 5 Why be a pluralist?

We have considered two distinct pluralist theories of reasons, which I have called 'pure F/D pluralism' and 'pure F/P pluralism'. But why would you want to be a pluralist in the first place?

[16] Consider his use of the expression 'the agent's normative reason for acting.' (Smith 1994, 131–32). Consider also the following: 'The distinction is that between psychological states that teleologically explain [i.e. motivating reasons] and considerations that justify [i.e. normative reasons]. The importance of making this distinction in this way becomes clear when we ask whether all actions must be done for reasons. For though this question gets answered resoundingly in the affirmative when reasons are understood to be motivating reasons… the question gets answered just as resoundingly in the negative when reasons are understood to be normative reasons.' (Smith 2004, 174–75)

## 5.1 Pluralism respects the 'two senses' intuition

Well, if it strikes you (as it strikes me) that it is true that there is a sense in which Sally runs for a reason and a sense in which she doesn't, or that there is a sense in which I have a reason to buy milk and a sense in which I don't, then you have already set the stage for a plural theory of reasons. Indeed, it seems to me, if you want your theory of reasons to respect this 'two senses' intuition (which you might, if it is an intuition you share) then you *have* to adopt a plural theory of reasons.

## 5.2 Pluralism may solve the problems that univocal theories face

A second appealing feature of pluralist theories of reasons is that they seem to provide one with a means of solving many of the problems faced by the univocal claims discussed in the previous chapters.

For instance, consider The Deliberate Action Problem discussed in § (II)4.2, which consisted of the following set of mutually inconsistent claims:

**The Deliberate Action Problem**

(F4) Sally ran deliberately.

(F3) If *A φs* deliberately then *A φs* for a reason.

(F7) If *A φs* for a reason then there was a reason, *p*, for *A* to *φ*.

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(F10) Sally's running was not at all worth doing.

According to a pluralist theory of reasons, this problem needs re-formulating in a manner that brings to light *which* sense of each reason expression the claim is being made about. So, denoting the different senses of each reason expression as 'reason$_A$' (favourist) and 'reason$_B$', (either deliberativist or psychologist – depending on one's theory), we can re-formulate this problem as the following set of mutually inconsistent claims:

**A re-formulation of The Deliberate Action Problem**[17]

(F4) Sally ran deliberately.

(F3') If *A φs* deliberately then *A φs* for a reason$_B$.

(UNI) *A φs* for a reason$_B$ if and only if *A φs* for a reason$_A$.

(F7') If *A φs* for a reason$_A$ then there was a reason$_A$, *p*, for *A* to *φ*.

(FAV') For any *p*, *p* is a reason$_A$ for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(F10) Sally's running was not at all worth doing.

By adding (UNI), this re-formulation makes a premise of the problem that is implicit in its original formulation, explicit: that 'the reason for which one acts' is a univocal expression. A pluralist theory of reasons then solves The Deliberate Action Problem by rejecting (UNI), which thus avoids the need to reject the other *prima facie* reasonable claims.

And, indeed, pluralist theories of reasons can solve most of the problems considered in previous chapters in a similar manner. However, they don't solve all of them, and pluralist theories face new problems of their own, as the next sections will show.

# 6 Pluralism is no panacea

Since I share the 'two senses' intuition, I think that our eventual theory of reasons ought to be pluralist. However, as I will argue in this section, *just* adopting a pluralist theory of reasons is not enough to solve the problems considered in the previous chapters. Firstly, 'conventional' pluralist theories (i.e. those made up of only favourist, psychologist or deliberativist claims) cannot solve all of the problems considered in previous chapters. And secondly, the pluralist solution to any given problem relies on there being an implicit univocality assumption in that problem – but since it is not at all clear that all of the problems considered in previous chapters implicitly include such an assumption, it is not at all clear that pluralism really provides us with a solution.

[17] This is only meant to be an indicative example of how the pluralist might respond; in particular, there are other ways to formulate this problem that change *which* reason expression the univocality assumption concerns. For example, compare (UNI) with (UNI*) in the following, alternative re-formulation of The Deliberate Action Problem:

(F4) Sally ran deliberately.
(F3') If *A φs* deliberately then *A φs* for a reason$_B$
(F7*) If *A φs* for a reason$_B$ then there was a reason$_B$, *p*, for *A* to *φ*.
(UNI*) For any *p, p* is a reason$_B$ for *A* to *φ* if and only if *p* is a reason$_A$ for *A* to *φ*.
(FAV') For any *p, p* is a reason$_A$ for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.
(F10) Sally's running was not at all worth doing.

## 6.1 Conventional pluralism cannot solve all the problems

There is a problem that a pluralist theory that only consists of favourist, psychologist or deliberativist claims,[18] i.e. a *conventional* pluralist theory, cannot solve – The Experiential Reasons Problem.

Recall that favourism, psychologism and deliberativism all face some form of The Experiential Reasons Problem because none of them are compatible with the *prima facie* reasonable claims that (i) Sally's hearing a bear-like sound could be her reason for running; or (ii) that my reading that my friend had won award could be my reason for congratulating her. That being so, any conventional pluralist theory will also face The Experiential Reasons Problem.

Furthermore, it is worth noting that since both favourism and deliberativism face some form of The Psychological Reasons Problem, any pluralist theory that only consists of favourist or deliberativist claims (such as pure F/D pluralism), will also face The Psychological Reasons Problem.

The point of these remarks is this: even if we adopt a (conventional) pluralist theory of reasons, that still will not help us solve all the problems considered in the previous chapters. If we are to do that, we need a new family of claims about reasons.

## 6.2 Just because you *could* doesn't mean you *can*

Secondly, we should be sceptical about the pluralist's approach to solving the problems considered; as Dancy puts it: 'one cannot resolve philosophical puzzlement in this way by multiplication of senses.' (2011, 351)

One way to interpret Dancy's remark is the insistence that you can't solve these problems by just postulating different senses that aren't actually 'there'. However, as Dustin Locke rightly points out, this is not what the pluralist intends:

> I am not suggesting that we can *resolve* philosophical puzzlement by *multiplication of senses*. Rather, I am claiming that senses are *already* multiple—the phrase 'S's reason [for acting]' already has two distinct senses. (Locke 2015, 218)

To re-interpret Locke's argument in my own terms: if you have the 'two senses' intuition it is *that* intuition that makes you think that reasons are plural, which *then* forms the basis for a plural theory of reasons. That is, it's not that you are multiplying senses in order to solve the problems, it's that (at least if you have the 'two senses' intuition) the senses *just are* multiple, and once you acknowledge that, many of the problems that a univocal theory faces can fall

[18] As opposed to the claims of some new family, not yet considered.

away. If the 'two senses' intuition is right, then, the pluralist can insist that our philosophical puzzlement has arisen only because we haven't recognised that the senses already are multiple.

However, a more nuanced interpretation of Dancy's objection is harder for the pluralist to avoid. I think that Dancy can reasonably be interpreted as having meant that even if the senses are multiple, the fact that invoking multiple senses of the relevant reason expressions *could* solve many of the problems considered in the previous chapters should not make us think that they *can* be solved that way (i.e. that that is the correct solution to them).

It might be that the *prima facie* reasonable claims in any given problem are really restricted to a *single* sense of the reason expressions involved – in which case the assumption of univocality does no work, and the problem returns. That is, for instance, The Deliberate Action Problem may actually be like this:

**Another re-formulation of The Deliberate Action Problem**

(F4) Sally ran deliberately.

(F3*) If *A φs* deliberately then *A φs* for a reason$_A$.

(UNI) *A φs* for a reason$_B$ if and only if *A φs* for a reason$_A$.

(F7') If *A φs* for a reason$_A$ then there was a reason$_A$, *p*, for *A* to *φ*.

(FAV') For any *p*, *p* is a reason$_A$ for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(F10) Sally's running was not at all worth doing.

In this construal, because all the *prima facie* reasonable claims are all about the same sense of reason, 'reason$_A$', the univocality assumption does no work. So, a pluralist theory of reasons would thus still have to find some claim, in addition to (UNI), to reject – the problem returns.

A pluralist solution to any given problem relies on that problem being the result of a conflation of different senses of the same reason expression – that is, it relies on a univocality assumption being a part of every problem. The difficulty for pluralist solutions to the problems considered is that it seems quite plausible that at least some of the problems considered *do not* result from a conflation of senses, and, in that case pluralism is no help in solving them.

So, again, it seems that if we do want to solve all of the problems considered in the previous chapters, we need a new family of claims about reasons.

# 7 A challenge for pluralism

It is generally[19] agreed that whenever we give an agent's reason for acting we explain the agent's action in a way that makes them seem rational.[20] For instance, saying that Sally's reason for running was that she believed that a bear was chasing her explains the fact that she ran in a way that makes her seem rational for running; and, likewise, saying that my reason for congratulating my friend was that she had won an award explains why I congratulated her in a way that makes me seem rational for having done so.

This observation is of some relevance to pluralism because it is seemingly not restricted to any particular sense of the 'agent's reason for acting' expression. That is, it seems as though, we may add a further *prima facie* reasonable claim to those already considered:

(S1) Whenever we give an agent's reason for acting, *whatever the sense of the expression used*, we explain their action in a way that makes them seem rational.

Why is this claim problematic for pluralism? Well, recall the following:

(F5) Sally's reason for running was, *inter alia*, that she believed a bear was chasing her.

(P1) My reason for congratulating my friend was, *inter alia*, that she had won an award.

Only pure F/P pluralism[21] can accommodate the truth of both (F5) and (P1); it does so by insisting that a different sense of the expression, 'the agent's reason for acting', is being invoked in each case (psychologistic in the former, favourist in the latter). This view is problematic because, if it were true, it would be hard to see why our respective reasons for acting can both be cited in an explanation of our actions that makes us each seem rational.

Consider: according to pure F/P pluralism, the relations between (i) Sally's belief and her action and (ii) the fact that my friend won an award and my action are different. And yet, giving either explains our respective actions in a way that makes us seem rational. So the pure F/P pluralist is seemingly forced to say that in spite of the reason-relations being different, by incredible coincidence, both Sally's and my reason for acting end up standing in a common relation to our respective actions – the relation in virtue of which giving the agent's reason for acting explains their action in a way that makes them seem rational. Pure F/P pluralism thus

[19] (E.g. Dancy 2000, 8; Stout 2009, 53; Gibbons 2010, 343; Broome 2013, 47)

[20] This is not to assume that it is the agent's reason that does the explaining: one can hold this view without holding that it is the agent's reason for acting that explains their action, as, for instance, Dancy (2014) does.

[21] Or less 'pure' variants of it, such as: (F/F, F/F, F/F, F/P).

faces an unenviable dilemma: either it accepts this highly implausible coincidence[22] or it rejects this new *prima facie* reasonable claim.[23]

This is a particular problem for F/P pluralism, but it is also a constraint on any pluralist theory of reasons. That is, any pluralist theory of reasons will have to provide some account of how (S1) could be true, or suffer the consequences of rejecting it. I will return to this point in the final chapter, when I set out my own pluralist theory of reasons.

# 8 Conclusion

I have suggested that while the 'two senses' intuition might provide us with some motivation for adopting a pluralist theory of reasons, we cannot rely on pluralism to solve the problems considered in the previous chapters. Instead, I have argued, we need a new family of claims about reasons. The aim of the remaining chapters is to advance and then defend such a family.

[22] The 'highly implausible co-incidence' being that giving either sense of the expression, independently, explains the agent's action in a way that makes them seem rational.
[23] I.e. (S1).

# (VI)

# A new family of claims about reasons

> *In which I set out a new family of claims about reasons, and introduce the major challenge to it. I define* 'pro tanto *rational' actions as actions that an agent takes to be, in some respect, worth doing. I set out a new family of claims about reasons,* explanatory rationalism*, which says that all practical reasons explain why the actions for which they are reasons are* pro tanto *rational. I introduce the major challenge for explanatory rationalism, The Explanatory Exclusion Problem, which argues that only features of an agent's psychology could explain either why they do something or why it was rational for them to do it. I set out the program for the forthcoming chapters.*

In § (I), I noted that most theories of reasons subscribe to one, if not several of the following claims: favourism about reasons to act, psychologism about the reasons for which we act and/or deliberativism about the reasons for which we act (see Table I-6). In §§ (II)-(IV), I showed that each of these accounts is inconsistent with several *prima facie* reasonable claims.

In § (V), I argued that, even if the *senses* of any given reason expression are plural, pluralism is no panacea: that is, we cannot just rely on the plurality of senses as the way to make those *prima facie* reasonable claims consistent with our theory of reasons. In short, I argued, we must look beyond favourism, psychologism and deliberativism: we need a new family of claims about reasons.

In this chapter I introduce a new family of claims about reasons: *explanatory rationalism*. According to explanatory rationalism, the fundamental reason-relation is that of explaining why an action is *pro tanto* rational; which is to say that all practical reasons explain why the actions for which they are reasons are *pro tanto* rational. Now, since explanatory rationalism rejects favourism about reasons to act, psychologism about the reasons for which we act and deliberativism about the reasons for which we act, it can solve all of the problems set out in §§ (II)-(IV). My aim, in the chapters that follow, is to set out and defend explanatory rationalism.

In this chapter I define what it is for an action to be *pro tanto* rational, I describe explanatory rationalism and I set out the challenges for it.

## 1 ***Pro tanto*** **rational action**

Fevzi is waiting to board a flight to Japan. It's early so he had to miss his morning swim. He looks forlornly out of the window, yearning to go swimming. He could abandon his trip and

have his swim but he thinks it would be better to stay and board the flight. Nonetheless, he still thinks that there is something to be said for going swimming. We can say this about Fevzi: he thinks that swimming is, in some respect, worth doing, but he thinks that boarding the flight is *all things considered* worth doing.

In contrast, Fevzi sees nothing of worth in aimlessly wandering around the airport. If he went for a wander he'd miss the flight without, at least by his lights, any good coming of it; he thinks that wandering aimlessly is, *in no respect*, worth doing.

There are clear differences between these three actions (flying, swimming, wandering) that are, I submit, of relevance to their rational standing. First, I will introduce some terminology to characterise those differences and then I will argue that what differentiates the three actions is relevant to their rational standing. In particular, I will suggest that even though it is not rational for Fevzi to go swimming, it is nonetheless more rational for him to go swimming than it is for him to wander aimlessly around the airport, and that we should have some terminology that reflects that.

## 1.1 Two kinds of rational action

The rational thing for Fevzi to do is to board his flight. If he were to go swimming we would say that he acted irrationally since, by his own lights, it was not *all things considered* worth doing. Doing one thing when you believe something else to be *all things considered* more worth doing is not a rational thing to do.

If we leave aside the question as to what makes an action worth doing,[1] I think we can (hopefully uncontroversially) characterise a familiar sense of rational action as follows:

> *Assumption* It is rational for *A* to *φ* if and only if *A* takes *φ*ing to be, all things considered, worth doing.[2]

This is an assumption, it is not a definition or an analysis of what it is for an action to be rational, nor is it a claim about why an action is rational (we will come to that shortly). This is meant to be a fairly bland assumption about what obtains when it is rational for some agent to act; it leaves unspecified all of the details that would actually furnish us with a theory of

[1] Depending on one's theory of rationality or 'motivation', saying that Fevzi takes getting on the plane to be, in some respect, *worth doing* could be cashed out by saying that he believes that by boarding the plane he will go to Japan and either that he wants to go to Japan or that he judges that going to Japan would be good, or right, or something of the sort. Nothing that I have to say is meant to express a commitment to either of these views. See § (II)3.2 for related caveats.

[2] For instance, this is, I think, in keeping with Parfit's (2011, 34) characterisation of rational action. Compare also: 'an agent is shown to be acting rationally if, as we might put it, he is shown to be trying to do what there is good reason to do, even if as a matter of fact he is quite mistaken on that front.' (Dancy 2004, 33)

rationality. I state it here only so that we may better understand the concept of *pro tanto* rationality, which I will introduce forthwith.

What separates Fevzi's boarding the flight from both his going swimming and his wandering aimlessly around the airport is that it is rational for him to do the former, but not the latter. However it is nonetheless common to distinguish acts like Fevzi's going swimming from acts like his wandering aimlessly, in spite of the fact that they are both irrational. While it would be plainly irrational of Fevzi to do the latter, if he were to go swimming that would be a typical example of an *akratic* action, or so-called 'weakness of will'.

I suggest that what separates akratic acts from 'plainly' irrational ones is of relevance to their rational standing. If Fevzi were to go swimming that would be markedly more rational than if he were to wander aimlessly around the airport; that is, there would be markedly more by way of rational intelligibility in his action if he swam than if he wandered – and that is so even if we concede that going swimming is nonetheless not a rational thing to do.

Now, if what separates akratic behaviour from what I have called 'plainly irrational' behaviour is of relevance to the rational standing of those actions, then, I suggest, we need an intermediate concept of rationality that allows us to recognise the former as somehow *more* rational than the other. To that end, let us define '*pro tanto* rational actions' thus:

> *Definition* It is <u>*pro tanto* rational</u> for *A* to *φ* if and only if *A* takes *φ*ing to be, in some respect, worth doing.[3]

This is a definition, not a claim.[4] I have used the term 'rational' because (for the reasons just outlined) I believe that an action's being *pro tanto* rational is of some relevance to the action's rational standing (a *pro tanto* rational action, is, I submit, in a familiar sense of the word more *rational* than an action that is not even *pro tanto* rational). If the reader balks at terminology that says that it is *to any extent* rational for Fevzi to go swimming, then please feel free to substitute some less objectionable term in its place (*mutatis mutandis* throughout this discussion).[5]

Lastly, in order to make the distinction between rational action and *pro tanto* rational action clear, I suggest that we can refer to the former as *all things considered* rational action, noting

[3] Daniel Whiting (2014, 5) uses the terminology of *pro tanto* rational action equivalently. Parfit (2011, 34) also indicates the distinction I have suggested when he distinguishes between actions that are 'less than fully rational' and actions that are 'irrational'.

[4] That is, I am just defining the use of the technical '*pro tanto* rational' predicate here.

[5] I believe that a *pro tanto* rational action is an action that on Smith's (1987) terminology the agent is motivated to do (note: the state of being motivated is defeasible, on Smith's account), so I could have talked in terms of 'motivation'. However, I have avoided the terminology of 'motivation' as I think the dangers of misunderstanding are even more pronounced there.

that this is the typical understanding of 'rational' action. So, for the sake of clarity, here is how the two concepts apply to Fevzi's possible actions:

| **Action** | *All things considered* rational? | *Pro tanto* rational? |
|---|---|---|
| Boarding his flight | ✓ | ✓[6] |
| Going swimming | × | ✓ |
| Wandering aimlessly | × | × |

**Table VI-1**: The ways in which Fevzi's actions are (or aren't) rational

Saying that going swimming is *pro tanto* rational whereas wandering aimlessly is not even *pro tanto* rational allows us to recognise that there is a difference in the rational standing of the two actions without impinging on the fact that boarding the flight is the only *really* rational thing for Fevzi to do.

## 2 Explanatory Rationalism

The family of reason claims I want to introduce says that reasons of any kind explain why the actions for which they are reasons are *pro tanto* rational. Because this family of claims emphasises the *explanatory* character of the reason-relation, and because a reason explains why an action is *rational*, I will call this family of claims 'explanatory rationalism'. Using the schema developed in § (I), we can represent explanatory rationalism as follows:

| **Reason expression** | **Explanatory rationalism** |
|---|---|
| For any *p, p* is a reason for *A* to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. |
| For any *p, p* is a reason for *A's φ*ing… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p, p* is a reason *A* has to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. |
| For any *p, p* is *A*'s reason for *φ*ing… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and explains (in the right way) why *A φ*'d. |

**Table VI-2:** Explanatory rationalism

According to explanatory rationalism, there are three kinds of reason: the expressions 'a reason for *A* to *φ*' and 'a reason *A* has to *φ*' pick out one kind, the expression 'a reason for *φ*ing' picks out another[7] and '*A*'s reason for *φ*ing' picks out a final kind.

[6] Any *all things considered* rational action is automatically a *pro tanto* rational action since if an agent takes an action to be *all things considered* worth doing they certainly take it to be, *in some respect*, worth doing.

In later chapters I will argue that explanatory rationalism avoids all of the problems set out in §§ (II)-(V) and that, indeed, explanatory rationalism characterises the *de facto* sense of each reason expression. However, in line with the two senses intuition, I will suggest that there is another sense to each reason expression, which is favourist. Before we get there, though, I will need to demonstrate how it is that explanatory rationalism could be true in the face of the major objection to it, which I will call 'The Explanatory Exclusion Problem'.

# 3 A problem for explanatory rationalism

Recall the following example from § (III): My friend has won a much-coveted award; I read about it in a newspaper so I call her up to congratulate her. I suggested that it was *prima facie* reasonable to claim that my reasons for congratulating my friend were, *inter alia*, that she won an award and that I read that she had won an award in the newspaper.

Now, according to explanatory rationalism, an agent's reason for acting both explains why it was *pro tanto* rational for them to do what they did and explains why they did it. Therefore, if explanatory rationalism is to be consistent with these *prima facie* reasonable claims about what my reasons for congratulating my friend were (as I intend it to be), the following must be true:

(R1) I congratulated my friend because she had won an award.

(R2) I congratulated my friend because I read that she had won an award.

(R3) It was *pro tanto* rational for me to congratulate my friend because she had won an award.

(R4) It was *pro tanto* rational for me to congratulate my friend because I read that she had won an award.

There is a familiar argument against the idea that facts about the external world can explain why we act (which (R1) supposes), which proceeds, broadly as follows: if, for instance, I believe that it is raining when it isn't, I will still take my umbrella – *because I believe that it is raining*. However, given that I need to believe that it is raining in order to take my umbrella[8], then even if I take my umbrella when it is raining, I must still take my umbrella *because I believe that it is raining*. But, if my belief that it is raining can explain my action whether it is true or false, then what explanatory work can the fact that it is raining do? None, the argument concludes – which must mean that only features of an agent's psychology can explain their actions – that

[7] As a result, something could be a reason for an agent to *φ* without being a reason for their *φ*ing. I discuss this point further in § (XVI)A.3.

[8] Assuming I see nothing else of worth in taking my umbrella.

is, (R1) is not literally true. The same argument applies to (R2), and, as I will show, can be generalised also to the explanation of why it is (*pro tanto*) rational for an agent to act; that is, it can be used to show that (R3) and (R4) are also false.

This argument, which I call 'The Explanatory Exclusion Problem', is the motivating argument for psychologism (and psychologism about the reasons for which we act, in particular) because it insists that only features of an agent's psychology can explain their actions, and therefore (given that an agent's reason for acting must explain their action), that only features of an agent's psychology could be amongst their reasons for acting.

## 4 An outline of what follows

To the extent that the argument considered in the previous section is right, explanatory rationalism cannot solve the problems considered in §§ (II)-(IV) – indeed, to the extent that that argument is right, explanatory rationalism just collapses into psychologism. The rest of this discussion is thus, for the most part, a response to the argument of the previous section.

In § (VII), I make some assumptions about the structural principles and logical properties of explanatory relations, which I will use throughout my discussion. In § (VIII), I use this framework to provide a formal construal of The Explanatory Exclusion Problem. In § (IX), I show how the Problem also precludes perceptual experiences from explaining why we act, thereby counting against (R2); and, further, how it applies to the explanation of why it is rational to act, thereby also counting against (R3) and (R4).

The *de facto* response to The Explanatory Exclusion Problem is to accept the conclusion and say that the *purported explanans* in (R1), i.e. the fact that my friend had won an award, is merely *elliptical* for the *real explanans*, which is the fact that I believed that my friend had won an award. Because this is a claim about how it is that a *normative reason*[9] to do some action could explain why someone does it, I call this 'the *elliptical* theory of normative reason explanation'.

An alternative and increasingly popular theory, which seeks to preserve the *bona fide* explanatory role of normative reasons in action explanation, rejects the conclusion of The Explanatory Exclusion Problem and says that the fact that my friend won an award explains my action *directly*. This is the *direct* theory of normative reason explanation.

In § (X), I set out both the *elliptical* and the *direct* theories of normative reason explanation, and I argue that they are each flawed in ways that should make us look for an alternative.

[9] I re-habilitate this terminology in § (X)1, understanding *normative reasons* as things that make actions, in some respect, worth doing; without associating them with any particular reason expression.

In §§ (XI) & (XII), I develop that alternative: the *indirect* theory of normative reason explanation. This theory argues that a normative reason explains an agent's action by explaining the features of the agent's psychology that, in turn, explain their action.

My argument for the indirect theory proceeds in two stages: first, in § (XI), I argue that we should reject the conclusion of The Explanatory Exclusion Problem because it is based on a false principle of explanation, which I call 'the exclusion principle'. The exclusion principle requires that only the most *proximal* explanations of some *explananda* explain it; but this is mistaken – most of the explanations we are interested in are, to some extent, *distal* explanations.

Then, in § (XII), I show how that insight helps inform the account of how normative reasons explain actions. I argue that normative reasons are *distal* explanations of our actions; they explain those features of our psychology that, in turn, explain our actions. I then show how the indirect theory can be used to show that both (R1) and (R2) are true.

In § (XIII), I suggest that the same reasoning accounts for the truth of (R3) and (R4). For instance, I argue that the fact that I read that my friend had won an award explains why it is *pro tanto* rational for me to congratulate her because it explains why I believed that she had won an award, which, in turn, explains why it was *pro tanto* rational for me to congratulate her.

However, I note, the transitivity of explanation fails on some occasions, in particular when the explanatory chain is a deviant causal chain. So, we need some account of why it is that the explanatory chain up to the fact that it is *pro tanto* rational for an agent to act is transitive if it isn't deviant, but isn't transitive if it is deviant. In §§ (XIV) & (XV), I provide such an account.

First, in § (XIV), I introduce the *mystery* relation. I argue that the mystery relation is a non-causal, transitive, explanatory relation that relates: the belief that *p* to some justification for it when that belief is justified; the belief that *p* to the fact that *p* when the belief that *p* is knowledgeable; a justification for the belief that *p* to the fact that *p* when that justification affords the opportunity for knowledge; and an action to some belief that explains why that action is *pro tanto* rational when that action is done intentionally.

In § (XV), I argue that the mystery relation is transitive with the non-causal explanatory relation involved in explaining why some action is rational, whereas merely causal relations are not. This allows me to distinguish between deviant cases (which lack the required chain of mystery relations) and non-deviant cases (which don't). This leads me to argue: (i) that because I know that my friend has won an award, the fact that she has won an award explains

why it is *pro tanto* rational for me to congratulate her (which is (R3)); and (ii) that because my belief that she had won an award is based on the fact that I read that she had won an award, the fact that I read that she had won an award also explains why it is *pro tanto* rational for me to congratulate her (which is (R4)).

Finally, in (XVI), I revisit explanatory rationalism and show how it avoids the problems faced by other theories. I conclude by setting out my preferred theory of reasons, *new pluralism*, which holds that explanatory rationalism tells us one sense of what it is to be a reason, whilst favourism tells us the other. I then show how new pluralism easily responds to the challenge for pluralist theories introduced in the previous chapters.

# (VII)

# We need to talk about explanation

> *In which I make some assumptions about explanation. I say what I mean by 'explains' and I state that I will talk as though* explananda *are facts* and explanantia *are propositions (whether or not they are). I distinguish two sorts of explanatory relation, 'fully explains' and 'partially explains', where a full explanation is sufficient for the truth of the fact that it explains and a partial explanation is an element (or subset) of a full explanation, and I make some assumptions about the logical properties of these relations. Lastly, I say that some fact is 'overexplained' just in case there are two genuinely different full explanations of that fact.*

In this chapter I make some assumptions about the structural principles and logical properties of explanatory relations. This is with a view to having a technical framework with which to more formally characterise The Explanatory Exclusion Problem, which will be discussed in the next chapter. While I think that the assumptions below are intuitive and hopefully uncontroversial,[1] I think that neither the thrust of The Explanatory Exclusion Problem nor the effectiveness of my solution to it depends on formalising explanatory relations in this manner.

## 1 What do I mean by 'explains'?

Here is what Broome has to say about the many meanings of 'explains':

> 'Explain' in common usage has various senses. In one of them, Darwin explained why evolution occurs. In another, *The Origin of Species* explains why evolution occurs. In a third, natural selection explains why it occurs. (Broome 2013, 48)

I, like Broome, wish to stick to his third sense of 'explains' in this discussion. That is, when I talk of something *explaining* something else, that which *explains* is meant to be the *explanans* and not a description of it or the one who describes it. Similarly, when I talk of something being an *explanation*[2] of something else, I mean to say that it is an *explanans* of that thing.[3]

This 'explains' relation, so understood, is, I suggest, the same relation as the one picked out by the 'because' or 'reason why' expressions, so I shall hereafter take them to be equivalent.

---

[1] I take several of these assumptions from Broome (2013).

[2] That said, in § (X), when I come to talk of normative reason explanation, I will mean something different (and more akin to the second sense).

[3] Kim suggests that there may be a fourth kind of 'explains' relation: 'The *explanans* relation relates propositions or statements; the *explanatory* relation relates events or facts in the world. The *explanatory* relation is an objective relation among events that, as we might say, "ground" the *explanans* relation, and constitutes its "objective correlate."' (Kim 1988, 226) The 'explains' relations as I will use it is thus what Kim calls 'the *explanans* relation', and notably not the objective relation that he takes to underpin it (what he calls 'the explanatory relation').

# 2 Ontological assumptions

## 2.1 On the ontology of *explananda*

Throughout this discussion I will assume, purely for expositional convenience, that *explananda* are facts. One consequence of this assumption is that, for instance, one can explain why it is raining if it is raining, but one cannot explain why it is raining if it is not raining. A second consequence is that when I talk of something explaining, for instance, an *action* or a *belief* this is not meant to imply anything about the ontology of those *explananda*.

## 2.2 On the ontology of *explanantia*

Throughout this discussion I will mostly treat *explanantia* as propositions (or sets of propositions).[4] Importantly, I do not assume, in my *exposition* that *explanantia* are true propositions: this is because I want to be able to recognise the possibility of non-factive theories within my discussion – although I will ultimately dismiss them, because I do think that explanation is factive.[5]

Explanatory rationalism implies that all practical reasons stand in explanatory relations to the rationality of the actions for which they are reasons. Thus, whatever sort of thing *explanantia* are, so too are reasons. This means, therefore, that I will be treating reasons as propositions also.

However, although I will talk as though *explanantia* and reasons are propositions, I am not arguing that they are. My theory is neither about *what explanantia are* nor is it about *what reasons are*; it is about the relation in which reasons stand to the actions for which they are reasons. So, if your preferred theory of explanation says that mental states or states of affairs can be *explanantia* then it is compatible with my theory that reasons could be mental states or states of affairs also; my point is not that reasons are propositions, my point is only that it is the ontology of the relata of explanatory relations that determines the ontology of reasons (because I think the reason-relation is ultimately an explanatory relation).[6]

[4] Again, while I might occasionally talk about (for instance) beliefs explaining things, that should not be read as implying that it is the belief (*qua* mental state) doing the explaining, as opposed to the fact that the agent has that belief.

[5] Given the assumption that explanation is factive (i.e. all *explanantia* are true), and that true propositions are facts, my treating *explanantia* as propositions, amounts to Broome's (2013, 48) convention of treating them as facts.

[6] This is not a trivial caveat: some are strongly of the view that mental states, and not facts about them, are an agent's reasons (e.g. Turri 2009), whilst others are seemingly of the view that it is only states of affairs (and not facts) that have the metaphysical 'oomph' needed to explain (e.g. Dancy 2000).

# 3 Full explanation and partial explanation

Suppose that Joanne's roof leaks and that it rained last night and that her carpet is now wet.[7] When I say that Joanne's carpet is wet because her roof leaks and she says that it is wet because it rained last night, although we each cite different facts by way of explanation of why the carpet is wet, it seems clear that, as Broome puts it, 'our explanations are not rivals, and we would not feel we were contradicting each other.' (2013, 49) Furthermore, although we each say that the facts we cite explain why the carpet is wet, neither fact is what we might call 'the whole story' of why the carpet is wet.

It seems natural to think that the reason why the explanations that Joanne and I give are neither rivals nor, individually, the whole story of why the carpet is wet, is because they are each part of, what Broome calls, 'one big explanation' – where that *one big explanation* is the whole story of why the carpet is wet. Supplementing this idea, Broome suggests that when giving an explanation we typically pick just some part of this big explanation and that 'which part we pick out will depend on our context: our background knowledge, our interests in the matter and so on.' (2013, 49)

I want to further regiment Broome's suggestion. I suggest that we call this 'one big explanation' a 'full explanation', and the elements (or subsets) of it 'partial explanations', so that all explanatory relations are either full or partial. In the following sections I will make some assumptions about the logical properties of the 'fully explains' and 'partially explains' relations.[8]

## 3.1 Full explanations

We could think of full explanations as complex propositions or sets of propositions; for ease of exposition I will use the latter, without, in doing so, intending any claim about the ontology of *explanantia*.

What differentiates a full explanation from a partial explanation, I suggest, is that a full explanation is sufficient for the truth of its *explanandum*, so that a set of propositions fully explains some fact only if it is sufficient for the truth of that fact. That is:

[7] This is Broome's (2013, 48) example.

[8] It is worth noting that Ruben (2004) makes much use of the distinction between full and partial explanation, which is, in many respects, similar to mine (though I am perhaps more prescriptive). Schnieder (2011) draws the same distinction between what he calls ' complete' and 'incomplete' explanations. See also Raz's (2009, 185–86) discussion of a 'complete reason why' for analogous remarks.

*Assumption* For any proposition *p*, some set, Δ, fully explains the fact that *p* only if Δ entails[9] *p*.[10]

However, just entailing the truth of some proposition is obviously not sufficient for fully explaining it (hence the above is not a bi-conditional). For instance, the fact that it is raining entails the fact that it is raining, but the fact that it is raining does not explain (fully or otherwise) the fact that it is raining.

To substantiate this point, we can assume that the 'fully explains' relation is[11]:

- *Irreflexive* – Nothing fully explains itself;
- *Asymmetric* – If *p* fully explains *q* then *q* does not also explain *p*; and
- *Non-monotonic* – If Δ is a full explanation of the fact that *p*, then adding some arbitrary proposition to Δ does not entail that the set so created is also a full explanation of the fact that *p*.

While the argument of subsequent chapters does not depend on explanatory relations having these properties, if at least some of them seem plausible then that should make it clear the extent to which mere entailment falls short of explanation (as entailment is a reflexive, non-symmetric, monotonic relation).

Assuming that full explanations are non-monotonic means that adding some arbitrary proposition into what is already a full explanation of some fact does not give you a further full explanation. I want to strengthen this assumption by requiring that full explanations never contain superfluous parts; that is, as Wedgwood puts it:

> The explanans… must not contain any irrelevant elements that could be stripped away without making it any less sufficient to produce the explanandum. (Wedgwood 2002, 363)

This is a strengthening of non-monotonicity because it requires not just that you cannot add an arbitrary proposition into a full explanation and still say that that enlarged set is a full

[9] In what way does a set of propositions that fully explains some other fact 'entail' it? I suggest that it logically entails it, so that full explanations necessitate their *explananda*. This may mean that a full explanation of the fact that *p* may include facts (such as facts about physical laws) that are seemingly extremely peripheral (though not irrelevant) to the question 'why is it the case that *p*?' However, it is important to note that I'm not saying that everything in a full explanation is the sort of thing that we would say 'explains' the *explanandum* – so one could accept that some fact is a part of a full explanation without accepting that it is the sort of thing that we would say 'explains' their action (this might be what we would call an 'enabling condition'). See further remarks in the next section.

[10] Cf. Schnieder says that a complete explanation is an explanation 'whose explanans is sufficient for the explanandum.' (2011, 450)

[11] Explanatory relations are commonly assumed to have these properties – see, for instance, Rosen's (2010) remarks about explanation (in general) in his discussion of grounding explanations. Although it is worth noting that there is some dissent on whether or not explanatory relations respect these properties (see e.g. Ruben 2004).

explanation, but that each element of the full explanation should be, in some sense, 'necessary' to it. This property is, in other areas, known as 'minimality'.[12] I characterise it formally as follows:

MINIMALITY For any proposition *p*, some set, Δ, fully explains the fact that *p* only if there is no Γ such that Γ fully explains the fact that *p* and Γ is a proper subset of Δ.[13]

Lastly, it is worth noting that while a full explanation is sufficient for its *explanandum*, at least some full explanations are not necessary for their *explananda.* For instance, suppose that the fact that I've just been to the gym together with the fact that I am always tired after I've been to the gym fully explains why I'm tired. The fact that it fully explains my tiredness in this instance clearly does not mean that whenever I am tired it is because I've been to the gym etc., which is to say that at least some full explanations are not necessary for their *explananda*.

## 3.2 Partial explanation

We can now define the concept of 'partial explanation' in terms of the concept of 'full explanation', as follows: any element of a full explanation of some fact is a *partial explanation* of that fact. That is:

*Definition* For any propositions, *p* and *q*, *p* partially explains the fact that *q* if and only if there is a set, Δ, such that Δ fully explains the fact that *q*, and *p* is an element of Δ.

For the sake of completeness, it is worth noting that a set can also be a partial explanation (if it is a subset of a full explanation):

*Definition* For any proposition, *p*, and set, Γ, Γ partially explains the fact that *p* if and only if there is a Δ such that Δ fully explains the fact that *p* and Γ is a subset of Δ.

Some house-keeping: Firstly, just because some proposition (or set) is a partial explanation of some fact, we need not say that it *explains* that fact. That is, while we might say that the fact that Joanne had carpet under the hole in her roof is a part of the full explanation of why her carpet is wet (it's is a necessary part of the sufficient condition) we might not want to say that that fact *explains* why her carpet is wet. Like Broome, I suggest that what determines whether

[12] See, for instance, Audi's (2012b, 699) characterisation of *minimality* in the context of grounding. My formalization is a transposition of his.

[13] See, for instance, Raz's (2009, 185–86) remarks about the importance of non-redundancy to explanations.

or not we say that some partial explanation *explains* some other is a matter of background knowledge and other features of the context.[14]

Secondly, however, I will assume that if something *explains* some fact then it is a partial explanation of that fact (note this is not merely restricted to what we *say* explains the fact, but what actually explains it). Although I struggle to see how, if one accepts the basic structure of full and partial explanation set out here, one could deny this, it is perhaps still worth stressing:

> *Assumption* For any propositions, *p* and *q*, if *p* explains the fact that *q* then *p* is a partial explanation of the fact that *q*.

Thirdly, the 'partially explains' relation, so understood, is a contingent relation (unlike the 'fully explains' relation); the fact that it rained last night only partially explains why Joanne's carpet is wet *given* other facts about the way the world is. However, even though that fact only partially explains why the carpet is wet *given* other facts, the 'partially explains' relation is nonetheless still between that fact and the fact that the carpet is wet.

Fourthly, and more trivially, it follows from the above that all full explanations are also partial explanations (since any full explanation is a subset of a full explanation).

Lastly, I will assume, for the sake of completeness, that the 'partially explains' relation is (like the 'fully explains' relation), irreflexive, asymmetric and non-monotonic.

## 3.3 Explaining why the carpet is wet

Returning to our example, then: the full explanation of the fact that Joanne's carpet is wet includes, *inter alia*, facts such as the fact that it rained last night, the fact that her roof leaks and the fact that there is carpet beneath her leaky roof. That full explanation entails that Joanne's carpet is wet.

Each of the members of that full explanation (i.e. the fact that it rained last night, etc.) is a partial explanation of why Joanne's carpet is wet. However, contextual and pragmatic considerations will determine whether or not we say, of any given partial explanation, that it *explains* why her carpet is wet.[15]

[14] This is a suggestion about the *practice* of saying that one thing explains another – I make no comment on whether or not all partial explanations explain actions in some non-context relative sense.

[15] Cf. 'A partial explanation may be good relative to one set of circumstances, but bad relative to another, in which interests, beliefs, or whatever differ.' (Ruben 2004, 22)

# 4 Overdetermination and overexplanation

Having thus characterised full and partial explanation, I now want to offer a characterisation of *overexplanation* in terms of these concepts, as it is of some relevance to The Explanatory Exclusion Problem.

## 4.1 Overdetermination

Something is said to be *overdetermined* if and only if there are two separate sets of conditions that are each individually sufficient for it to obtain *and* those conditions *determine* that it obtain.[16] Different kinds of determination relation yield different sorts of overdetermination. Causal overdetermination (and whether or not it is possible) is probably the most hotly debated kind of overdetermination – here is a representative characterisation:

> Suppose that a certain event, in virtue of its mental property, causes a physical event. The causal closure of the physical domain says that this physical event must also have a physical cause. We may assume that this physical cause, in virtue of its physical property, causes the physical event… Could it be that the mental cause and the physical cause are each an *independent sufficient* cause of the physical effect? The suggestion then is that the physical effect is *overdetermined.* So if the physical cause hadn't occurred, the mental cause by itself would have caused the effect. (Kim 1993, 280–81)

The principle is something like this: something is causally overdetermined if you could take one of its causes away and it would still obtain (or occur, or exist or what have you). A classic example: two vandals each throw rocks that simultaneously strike and break a window. The breaking of the window is seemingly causally overdetermined because either rock-throwing would have been sufficient to break the window. It's beyond the scope of this discussion to get into what that means for the causal status of either rock-throwing.

## 4.2 Overexplanation

Here is how I propose to characterise *overexplanation*:

> *Definition* For any proposition *p*, the fact that *p* is overexplained if and only if there are (at least) two *genuinely different* full explanations of the fact that *p*.

Is there a difference between overexplanation and overdetermination? If you are an explanatory realist (so that explanatory relations [17] are underpinned by ontological

[16] I stress the latter conjunct as there being merely two sets of conditions that entail some fact is not sufficient for the fact to be over-determined (if it were then, arguably, everything would be overdetermined, given the reflexivity of the entailment relation, and the claim that everything is determined in some more metaphysically significant sense than entailment).

[17] Of the kind I have in mind (see fn. 3 of this chapter for clarification).

determination relations[18]) and you hold that explanations are *genuinely different* only if they are *independent*,[19] then you will probably think that overdetermination and overexplanation are the same thing.

However, since I want to allow for the possibility of rejecting either explanatory realism *or* that two explanations are genuinely different only if they are independent, I distinguish overexplanation from overdetermination. Distinguishing them in this manner does not preclude the possibility of there being the same; that depends only on how one defines the *definiens* of overexplanation.

What does it mean for two full explanations to be genuinely different? I suggest it is that they should explain the *explanandum* in different ways. Why call them 'genuinely different' and not merely 'different' explanations? Because I want to allow for the possibility that non-identical full explanations (i.e. 'different') full explanations may nonetheless explain some *explanandum* in the same way (i.e. without being 'genuinely different').

## 4.3 Benign overexplanation

It is widely believed that genuine causal overdetermination is rare (if it is even possible).[20] Assuming that causal determination relations underpin causal explanatory relations, causal overexplanation is presumably equally rare.

However, *bona fide* cases of non-causal overexplanation abound. For instance, recall that I said that swimming will both help me sleep better and improve my mood. Swimming is then, for me, in some respect, worth doing partly because it will help me sleep better and partly because it will improve my mood. The explanatory relations involved here are, I suggest, clearly not causal.

Now consider this: swimming would still be, in some respect, worth doing even if it wouldn't help me sleep better because it would still improve my mood. Conversely, if it weren't the case that swimming would improve my mood, we could still say that it was, in some respect, worth doing because it would help me sleep better.

[18] Cf.: 'According to "explanatory realism," when something is correctly invoked as an explanation of another thing, the explanatory relation must be grounded in some objective relation of dependence or determination holding for the *explanans* and the *explanandum*.' (Kim 1993, xii)
[19] That is, in whatever sense you take an overdetermining cause/factor to be *independent*.
[20] Kim (1993, 280) describes the idea that there could be systematic causal overdetermination as 'absurd'.

In contrast, if it hadn't rained last night then, *ceteris paribus*, the fact that Joanne's roof leaks would not continue to partially explain the fact that her carpet was wet (her carpet would not have been wet, if her roof didn't leak).

What differentiates these two cases is seemingly this: while the fact that it rained last night and the fact that her roof leaks are part of the *same* full explanation of why her carpet is wet, the fact that swimming would help me sleep better and the fact that it would improve my mood are part of two *genuinely different* full explanations of why swimming would be, in some respect, worth doing – they explain the *explanandum* in different ways. Thus, since there are two genuinely different full explanations of why swimming is, in some respect, worth doing, it is *overexplained*. Whereas, in contrast, since there are not two genuinely different explanations of why Joanne's carpet is wet, the fact that her carpet is wet is not *overexplained*.

The characterisation of overexplanation hangs on what it is for two explanations to be genuinely different. I have given what I take to be an unambiguous example here, but I will return to what makes explanations genuinely different in the next chapter (see § (VIII)3.2).

# 5 Summary

I have said that there are two kinds of explanatory relation: full and partial, and I've made some assumptions about the logical properties of each. I then provided an analysis of *overexplanation* in terms of full explanation. In the next chapter I will put these concepts to work in a characterisation of the main challenge to my theory: The Explanatory Exclusion Problem.

# (VIII)

# The Explanatory Exclusion Problem

*In which I set out The Explanatory Exclusion Problem, which is, in some form or another, the motivating argument for psychologistic theories of reasons. I provide a formal construal of the Problem, showing how it results from two seemingly trivial claims about what explains an agent's action when they act from error and from ignorance together with five seemingly plausible principles of explanation. I show how the Problem implies that I did not congratulate my friend because she had won an award, but only because I thought she had.*

My friend has won a much-coveted award; I read about it in a newspaper so I call her up and congratulate her. Did I congratulate her because she had won an award, or just because I thought she had?

There is a well-established response to this question that proceeds, broadly, along the following lines: if she hadn't won an award but I had believed that she had, then I would still have congratulated her, and I would have congratulated her *because* I believed that she had won an award. Conversely, if I hadn't believed that she had won the award then even if she had won it I would not have congratulated her. So, if I hadn't believed that she had won the award, the fact that she had won the award could not have explained why I congratulated her (since I wouldn't have).

So, it seems as though what matters to the explanation of why I congratulated my friend is the fact that I believed that she had won an award and not the fact that she had won an award; that is, I did not really congratulate my friend because she had won an award, but only because I thought she had.

An argument along these lines is what typically motivates the view that facts about the world cannot explain an agent's action and that, therefore, an agent's reason for acting must be a feature of their psychology.[1] Indeed, this line of reasoning is the motivating argument for what I called 'psychologism about the reasons for which we act'.[2] So, for those who want to reject that form of psychologism (as I do), this argument is the one to beat.

In what follows I offer a formal construal of the argument as an argument about what explains an agent's action (i.e. as an argument that is only indirectly about what their reason for acting is). I will use the concepts introduced in the previous chapter to formally characterise the three

[1] Since an agent's reason for acting always explains their action (see § (IV)1.2).
[2] See Table I-4.

components of what I will call 'The Explanatory Exclusion Problem', which are: the argument from false belief[3]; the argument from impotent facts[4]; and a principle of explanation that I call 'the exclusion principle'[5]. In particular, on my construal, the conclusion of The Explanatory Exclusion Problem can be arrived at from two seemingly trivial claims about what explains an agent's action when they act in error or ignorance, together with five seemingly plausible principles of explanation.

My intent in formalising The Explanatory Exclusion Problem is three-fold: firstly, doing so will help me demonstrate, in § (IX), that the Problem applies also to the facts on which an agent's beliefs are based[6] and to the explanation of why it is rational for someone to do something.[7] Secondly, formalising the Problem will help me to discriminate more easily between different responses to it. And, thirdly, a formal construal allows me to identify more precisely where The Explanatory Exclusion Problem goes wrong.

I should note that other construals of this argument are possible, and mine is by no means definitive (though I hope it is illuminating). However, I do not think one can construe the overall problem in a way that makes it immune to my eventual response; that is, I do not think that my response to the overall problem hangs on formalising it in the way that I do.

Lastly, we should also be clear that my discussion here is strictly about what explains an agent's action, and not what their reason for acting is. The Explanatory Exclusion Problem only bears upon what an agent's reason for acting could be to the extent that we assume that an agent's reason for acting always explains their action (although this is a widely held assumption - see § (IV)1.2).

# 1 An overview

Recall that I said that if explanatory rationalism is to be consistent with the *prima facie* reasonable claims set out in §§ (II)-(IV), the following must be true:

(R1) I congratulated my friend because she had won an award.

The Explanatory Exclusion Problem provides the following argument against this claim:

[3] This is Stout's (1996, 2009) name for this argument.
[4] Stout (1996) discusses what he calls, 'The Argument from the Impotence Unrepresented Facts'; while his argument is in some respects similar to mine, the premises and conclusions of our arguments are sufficiently different as to make them different arguments.
[5] This principle share's some similarities with Kim's principle of causal exclusion, but, as I shall note, differs from it in particular respects that shall turn out to be critical to this discussion.
[6] And not merely facts about the external world.
[7] And not merely to the explanation of why they do it.

**The Explanatory Exclusion Problem for (R1)**

| | |
|---|---|
| *Premise 1* | There is a full explanation of why I congratulated my friend such that the fact that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation.[8] |
| EXCLUSION | For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation[9], then *p* does not partially explain *q.* |
| *Conclusion 1* | The fact that my friend had won an award does not explain why I congratulated her. |

The focus of this chapter is on setting out the argument for Premise 1 and for EXCLUSION. In § (IX) I will show how, by altering Premise 1, The Explanatory Exclusion Problem can also provide arguments against (R2), (R3) and (R4).

## 1.1 The argument for Premise 1

Premise 1, as I will demonstrate, follows from the conclusions of two other arguments: *the argument from false belief* and *the argument from impotent facts*.

The conclusion of *the argument from false belief* is that we can give a full explanation of why I congratulated my friend without mentioning the fact that she won an award; so long as we note that I believed that she had won an award. I show how this conclusion can be arrived at from a seemingly trivial claim about what would have explained my action had my belief been false together with three seemingly plausible principles of explanation.

The conclusion of *the argument from impotent facts* is that the fact that my friend had won an award is not part of a genuinely different explanation of why I congratulated her from the fact that I believed that she had. I show how this conclusion can be arrived inferred from (i) a seemingly trivial claim about what the explanatory power of the fact that my friend won an award would have been if I hadn't believed that she had; together with (ii) another plausible principle of explanation.

Premise 1 can then be inferred from the conclusions of these two arguments.

## 1.2 The argument for EXCLUSION

The argument for EXCLUSION is more straightforward. In short: if some fact is not part of a full explanation of some *explanandum* and it is not part of a genuinely different explanation of

[8] That is, a genuinely different explanation of why I congratulated my friend… I omit this qualification throughout, for brevity.

[9] That is, a genuinely different explanation of why *q*… Again, I omit this qualification throughout, for brevity.

that *explanandum*, then seemingly, by the law of the excluded middle, it is not part of any full explanation of that *explanandum*, which means that it does not explain it.

## 1.3 What's next

In what follows I set out the argument from false belief and the argument from impotent facts. I then show the conclusions of these two arguments yield Premise 1. I then set out the argument for EXCLUSION. I then show how Premise 1, together with EXCLUSION, yields Conclusion 1 (which is the denial of (R1)).

For reference, the figure below provides an overview of the structure of this discussion.

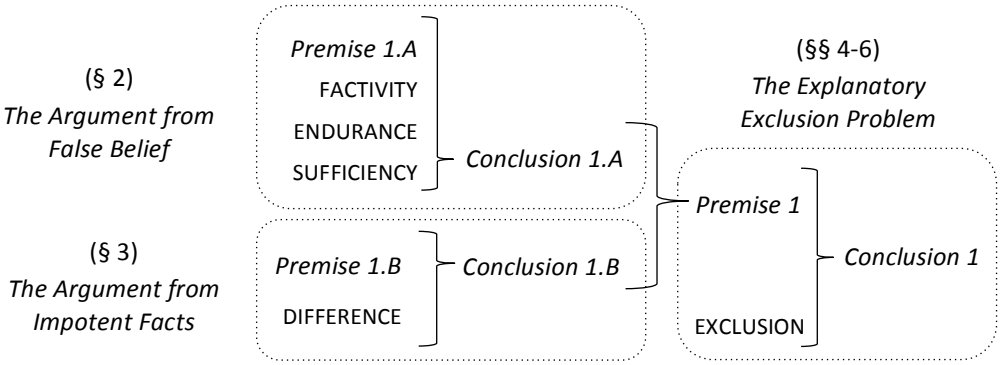


**Figure VIII-1:** The argument for The Explanatory Exclusion Problem

# 2 The Argument from False Belief

Dancy provides a concise summary of the argument from false belief, as it realtes to an agent's reason for acting:

> [Consider] the case where things are not as the agent conceives them to be. Surely, in such a case, we cannot say that his reason for acting as he did was that *p*. We have to say that his reason for acting was that he believed that *p*. Accepting this for the case where the relevant belief is false, then, we might still hope that 'that *p*' can indeed be the explanation of the action where it is the case that *p*, but that where it is not the case that *p* the explanation can only be 'that he believed that *p*'. But, as Bernard Williams puts it, the true-false distinction should not be allowed to affect the form of the relevant explanation. Supposing, therefore, that our explanation should take the same form whether it is or is not the case that *p*, and having already accepted that the correct explanation in cases where it is not the case that p is 'that he believed that *p*', we are driven to say the same where the relevant belief is true rather than false. (Dancy 2000, 121)

Adapting Dancy's argument to my present concern about what *explains* an agent's action (and not just their reason for acting), the main conclusion of this line of reasoning seems to be that we can always explain an agent's action in terms of their beliefs, without reference to the truth

or falsity of that belief. This conclusion may seem obvious, however, since some deny it,[10] it is worth being explicit about the argument for it.

In what follows I provide a formal construal of this argument. I take my construal to be plausible and, I hope, informative; however, other construals are available (e.g. Stout 2009). The conclusion of the argument from false belief, in the case of my friend's award, is as follows:

> *Conclusion 1a* There is a full explanation of why I congratulated my friend that includes the fact that I believed that she had won an award but not the fact that she had won an award.

This conclusion is the first part of the argument for Premise 1. In § 3, I present the argument from impotent facts, which is the second part of the argument for Premise 1.

## 2.1 Acting on false beliefs

The argument from false belief starts, predictably, with an observation about what happens when an agent acts on a false belief. Consider: even if my belief had been false, I would still have congratulated my friend, and I would have congratulated her *because* I believed that she had won an award.[11] Now, to say that I would have congratulated her because I believed that she had won an award is to say that the former partially explains the latter.

Moreover, I suggest, when I congratulate her because of my false belief that she won an award, the fact that she did *not* win an award is not part of any explanation of why I congratulated her.[12]

This much should be undeniable. Thus:

> *Premise 1a* If, *ceteris paribus*, my friend had not won an award (but I still believed that she had[13]), then (i) the fact that I believed that she had won an award would have partially explained why I congratulated her; and (ii) the fact that she had not won an award would not have partially explained why I congratulated her.

## 2.2 The Factivity Principle

Recalling the discussion of § (IV)1.3, where we noted that the factivity of explanation is a *prima facie* reasonable and widely held principle of explanation, let us just assume the following:

---

[10] See the discussion of exclusive disjunctivist theories in § (X)A.4.

[11] Compare: why did Sally run? Because she thought a bear was chasing her.

[12] Perhaps, one could craft a weird example in which it was, but that is not my example. Likewise, compare: the fact that a bear *wasn't* chasing Sally does not explain why she ran. The relevance of this qualification will become clear as the discussion proceeds.

[13] This remark is parenthetical because it is already implied by the *ceteris paribus* condition.

FACTIVITY For any propositions *p* and *q*, if *p* partially explains the fact that *q* then *p* is the case.

This means that when Sally runs because she mistakenly believes that a bear is chasing her the false proposition *that a bear was chasing her* does not explain why she ran. Likewise, if my belief that my friend had won an award had been false, the false proposition *that my friend had won an award* would not explain why I congratulated her.

## 2.3 Two more principles of explanation

Consider the following, much-discussed remark:

> The difference between false and true beliefs on the agent's part cannot alter the form of the explanation which will be appropriate to his action. (Williams 1981, 102)

Here is what I take to be an implication of what Williams is saying, in my terminology: the same full explanation of one's action is available whether one's belief is true or false.[14] While this claim may seem obvious, there are those who deny it, and who do so for different reasons, so it is worth considering it in more detail.

In what follows, I want to show how this claim can be motivated by two principles of explanation, *the endurance principle* and *the sufficiency principle*, that are both general (i.e. apply beyond the explanation of action) and plausible.

### 2.3.1 The Endurance Principle

The following remark, I suggest, relies on the endurance principle:

> When an agent acts on false beliefs, we cannot explain the action in terms of the facts but only in terms of those beliefs – there is only an internalist explanation of their action. *But even when the beliefs are true that same internalist explanation works.* (Stout 1996, 24 emphasis added)

Stout suggests that since we can explain an agent's actions in terms of the facts about what they believed when their belief was false, we can likewise explain their action in terms of the facts about what they believed when their belief is true. In which case, given the fact that Sally believed that a bear was chasing her (partially) explains why she ran when her belief was false, if, *ceteris paribus*, her belief had been true, the fact that she believed that a bear was chasing her would still have (partially) explained why she ran. But why should this be the case? Why does it follow that what explains in the false belief case also explains in the true belief case?

Let's start by noticing this: when Sally runs because she mistakenly thought that a bear was chasing her, we know, from FACTIVITY, that the false proposition *that a bear was chasing her*

[14] Williams's remark, I think, goes further than this – I think that, given that explanation is factive, there is a reading of his remark on which it just is the conclusion of the argument from false belief.

(i.e. the content of Sally's belief) could not partially explain why Sally ran. Furthermore, the fact that a bear was *not* chasing Sally is also not part of any full explanation of why she ran – and, in particular, it is not part of the *same* full explanation as the fact that she believed that a bear was chasing her.[15] So, there seems to be a clear sense in which, when Sally's belief is false, the proposition *that a bear is chasing her* is irrelevant to the *explanation* of why she ran.[16]

Now consider what would have happened if, *ceteris paribus*, Sally's belief had been true rather than false? This would have happened: a proposition that was irrelevant to the explanation of her action (i.e. *that a bear was chasing her*) would have gone from being false to being true. But why would that proposition's suddenly becoming true affect the pre-existing partial explanation relations if it was irrelevant to those partial explanations when it was false? That is, why would Sally's belief that a bear was chasing her stop explaining her action just because a false proposition that was irrelevant to that explanation suddenly became true.[17] Seemingly, it would not.

To the extent that this line of reasoning is persuasive, I suggest that that is just because it accords with a more general principle of explanation – namely that if, *ceteris paribus*, a proposition that is irrelevant to some explanation of some *explanandum* (i.e. neither it nor its negation is part of that full explanation of that *explanandum*) when it is false suddenly becomes true, then that does not stop anything that partially explained that *explanandum* when that proposition was false from continuing to explain it when it is true. That is, the partial explanation relations between facts *endure* when the truth-value of a proposition that is irrelevant to them changes. Thus:

[15] Again, one could craft a weird example in which it was, but that is not my example.

[16] This does not mean that the proposition is irrelevant in all senses of relevance – for instance, it is certainly something she took to make running worth doing; the point is just that from a particular explanatory perspective, it is seemingly irrelevant. It is also worth stressing this: just because the proposition *that a bear is chasing her* is irrelevant when it is false, does not mean that it is irrelevant when it is true – a proposition is irrelevant to some partial explanation of some fact only in some particular instance and only in so far as neither it nor its negation are part of the same full explanation of that fact as that partial explanation.

[17] Consider: even if the fact that a bear was chasing her is relevant to the explanation of her action *when true* that does not show that its becoming true would destroy pre-existing partial explanation relations, despite its prior irrelevance.

ENDURANCE For any propositions *p, q* and *r*, the following holds: Suppose that *q* partially explains the fact that *r* when it is not the case that *p.* Suppose further that neither *p* nor *not p* is part of the same explanation of *r* as *q*. Then, if, *ceteris paribus*, it were the case that *p*, *q* would still partially explain the fact that *r*.[18]

Although this claim is long-winded, I take the principle to be ultimately intuitive. However, in the event that it is not clear, the next section discusses some examples and a failed counterexample.

### 2.3.2 Some examples of the endurance principle

The reader who is already comfortable with ENDURANCE may skip these examples.

**Example 1:** An uncontroversial example: Suppose that a teacher is performing a science experiment for her students, although one of the students, Nathan, is absent. She heats a metal rod and it expands. The fact that the rod was heated partially explains why it expanded, and the false proposition *that Nathan is present* is irrelevant to that explanation of why it expanded.[19] Moreover, *ceteris paribus*, had Nathan been present, the fact that the rod was heated would still partially explain why it expanded.

**Example 2:** Of course, Nathan's being present is rarely likely to feature in an explanation of why the rod expanded, so here is another example, in which the proposition that is irrelevant when false, explains when true. Consider the case in which swimming will help me sleep better but won't improve my mood. We already established[20] that the fact that swimming will help me sleep better partially explains why swimming is, for me, in some respect, worth doing.[21] Moreover, I suggest, neither the false proposition *that swimming will improve my mood*, nor the fact that it won't improve my mood explain why it is, in some respect, worth doing – so the proposition *that swimming will improve my mood* is irrelevant to the full explanation of why swimming is, in some respect, worth doing.

However, if, *ceteris paribus*, swimming would improve my mood then it would partially explain why swimming was, in some respect, worth doing. But just because, once it obtains, the fact that swimming would improve my mood *starts* explaining why swimming would be, in some respect, worth doing, doesn't mean that the fact that swimming will help me sleep better

[18] More formally: For any propositions *p, q* and *r*, if, when it is not the case that *p, q* partially explains the fact that *r* then, provided that neither *p* nor *not p* is part of the same explanation of *r* as *q*, if, *ceteris paribus*, it were the case that *p*, then *q* would still partially explain the fact that *r*.

[19] Neither it nor its negation is part of the full explanation (of which the fact that the rod was heated is part) of why the rod expanded.

[20] See § (VII)4.3.

[21] Given that I want to sleep better, or judge it good and what have you.

*stops* explaining why it is, in some respect, worth doing. All that happens now is that they *both* explain.

**Example 3:** A would-be counterexample: suppose that Tom throws a rock at a window and it breaks, while Susie stands and watches. The fact that Tom threw a rock at the window (partially) explains why it broke. Seemingly, neither the false proposition *that Susie threw a rock at the window* nor the fact that she didn't throw a rock explains why the window broke: the proposition *that Susie threw a rock* is irrelevant to the full explanation of why the window broke.

However, suppose that Susie had thrown a rock, and she had thrown it before Tom, and her rock had broken the window. In which case Tom's rock-throwing would seemingly cease to explain why the window broke (his rock would have sailed through the empty space where the window used to be). So, *contra* the endurance principle, when an irrelevant proposition goes from true to false that *can* stop something from partially explaining the *explanandum*.

This counterexample fails because it violates the '*ceteris paribus*' condition in ENDURANCE. In making it the case that Susie threw a rock at the window *and* it broke we are not changing the truth-value of *only* irrelevant propositions, but of *relevant* propositions as well. For instance, the fact that the window was intact before Tom's rock hit it is a part of the same full explanation of why the window broke as the fact that Tom threw the rock.[22] So of course the fact that Tom threw the rock stops explaining why the window broke – by adding in Sally's rock-throwing we have taken away an element of the full explanation that Tom's rock-throwing was part of, so everything in it stops explaining.

In contrast, suppose we honour the '*ceteris paribus*' condition and change only that which is not part of the same full explanation of the window's breaking as Tom's rock throwing. So, let's suppose that Susie's rock harmlessly bounces off the window. In this circumstance it should still be clear (given that all other things are equal) that Tom's rock breaks the window and Tom's rock-throwing explains why the window broke. ENDURANCE perseveres.

### 2.3.3 The Sufficiency Principle

Now: another principle of explanation – the sufficiency principle. Recall that I interpreted Williams's remark as the claim that the same full explanation of an agent's action is available whether their belief is true or false. We need more than just ENDURANCE to reach that conclusion.

[22] It is a necessary part of that full explanation because if the window hadn't been intact, it wouldn't have been broken by Tom's rock.

Consider: there is a set of partial explanations of why Sally ran (when her belief was false) that is a full explanation of why she ran. What ENDURANCE tells us is that that which partially explains her action when her belief is false would also partially explain if, *ceteris paribus*, her belief were true. So, the same set of partial explanations that fully explained her action when her belief was false would still be a set of partial explanations of her action if her belief were true. However, we need to add some further requirement to guarantee that that set is still a full explanation when her belief is true (as opposed to being *merely* a set of partial explanations i.e. an incomplete full explanation). What is that requirement?

It is this: whatever suffices to explain an agent's action when an agent's belief is false likewise suffices to explain it when their belief is true. And, again, I suggest that this is just a consequence of an intuitively plausible, and more general principle of explanation – namely that if some set of partial explanations suffices to explain (i.e. is a full explanation of) some fact in some situation, then whenever those partial explanations all explain that fact, they will suffice to explain it. Thus:

> SUFFICIENCY For any proposition *q*, and any set, Δ, if Δ is a full explanation of the fact that *q* in some circumstance, then, in any circumstance in which all the elements of Δ partially explain the fact that *q*, Δ fully explains the fact that *q*.[23]

### 2.3.4 Combining the Endurance Principle and the Sufficiency Principle

The endurance principle and the sufficiency principle provide us with the conclusion that the same full explanation of an agent's action is available whether their belief is true or false. How? By ensuring that just changing the truth-value of some proposition that is outside of a full explanation of some *explanandum* (i.e. which is irrelevant to the explanation of that *explanandum*) cannot affect whether or not that full explanation is available. So, if you think that neither the proposition *that a bear is chasing her*, nor the fact that a bear is not chasing her, are parts of the full explanation of why Sally ran (when her belief was false), then that same full explanation will be available, *ceteris paribus*, regardless of whether or not a bear is chasing her.[24]

---

[23] This may seem close to MINIMALITY, but it is a very distinct claim. However, the conjunction of the two yields what we might call *counterfactual minimality*, which can be defined as follows:

> C-MINIMALITY For any proposition *p*, some set, Δ, fully explains the fact that *p* only if there is no Γ such that Γ is a proper subset of Δ and if, *ceteris paribus*, Δ had not existed but Γ had, then Γ would fully explain the fact that *p*.

I suspect that an intuitive commitment to C-MINIMALITY, resulting from an implicit commitment to both MINIMALITY and SUFFICIENCY as principles of explanation, helps motivates the argument from false belief.

[24] Note that non-factivists (e.g. Dancy 2000), who think that explanation can be non-factive, would say that *that a bear is chasing her* (*qua* the content of her belief) is part of the full explanation of why Sally

This may seem elaborate, but distinguishing these two general explanatory principles is essential not only to discriminating between different rejections of The Explanatory Exclusion Problem,[25] but also to clearly demonstrating why it is wrong.

## 2.4 Concluding the argument from false belief

Here, then, is my construal of the argument from false belief: if, *ceteris paribus*, my friend had not won an award then the fact that I believed that she had won an award would have been a part of a full explanation of why I congratulated her (from Premise 1a). However, neither the fact that she had *not* won an award nor the false proposition *that she had won an award* would have been part of that explanation (from Premise 1a and FACTIVITY).

We should make two observations from these remarks about what would have been the case if my belief had been false: firstly, there would have been a full explanation of why I congratulated my friend that would have included the fact that I believed that she had won an award but not the (false) proposition *that she had won an award.* Secondly, the (false) proposition *that she had won an award* would have been irrelevant to the explanation of my action.

From the second observation, we can conclude (from ENDURANCE) that whatever would have partially explained my action if my belief had been false must also partially explain it when, *ceteris paribus*, my belief is true. So, since all the elements of what would have fully explained my action had my belief been false also (partially) explain my action when my belief is true, we can infer (from SUFFICIENCY) that that set of partial explanations must likewise fully explain my action when my belief is true.

Now recall the first observation: had my belief been false there would have been a full explanation of my action that included the fact that I believed that my friend had won an award but not the (false) proposition that she had won an award. But given that the same full explanation is available whether my belief is true or false, that must mean that even when my belief is true, there is a full explanation of my action that includes the fact that I believed that my friend had won an award but not the fact that she had won an award.

---

ran. This, however, would not stop them from agreeing with the claim that the same full explanation of Sally's action is available whether her belief is true or false (indeed, the desire to agree with Williams' claim is a part of what persuades Dancy to adopt non-factivism).

[25] In particular: those whom I call 'exclusive disjunctivists' (e.g. Collins 1997; Stoutland 1998) reject ENDURANCE, whereas those whom I call 'supplementarists' (possibly Alvarez 2010) reject SUFFICIENCY. See the Appendix to § (X) for further discussion.

Formally:

**The Argument From False Belief**

| | |
|---|---|
| *Premise 1a* | If, *ceteris paribus*, my friend had not won an award (but I still believed that she had), then (i) the fact that I believed that she had won an award would have partially explained why I congratulated her; and (ii) the fact that she had not won an award would not have partially explained why I congratulated her. |
| FACTIVITY | For any propositions *p* and *q*, if *p* partially explains the fact that *q* then *p* is the case. |
| ENDURANCE | For any propositions *p, q* and *r*, the following holds: Suppose that *q* partially explains the fact that *r* when it is not the case that *p.* Suppose further that neither *p* nor *not p* is part of the same explanation of *r* as *q*. Then, if, *ceteris paribus*, it were the case that *p*, *q* would still partially explain the fact that *r*. |
| SUFFICIENCY | For any proposition *q*, and any set, Δ, if Δ is a full explanation of the fact that *q* in some circumstance, then, in any circumstance in which all the elements of Δ partially explain the fact that *q*, Δ fully explains the fact that *q*. |
| *Conclusion 1a* | There is a full explanation of why I congratulated my friend that includes the fact that I believed that she had won an award but not the fact that she had won an award. |

The conclusion of the argument from false belief provides the first part of the argument for Premise 1. The argument from impotent facts, which is the subject of the next section, provides the second part of that argument.

# 3 The Argument from Impotent Facts

The argument from impotent facts is, mercifully, simpler. It is, in some sense, summarised in the following remark:

> Whenever the agent acts in light of the fact that *p* [i.e. because *p*], the agent must take it that *p*, and I understand this sort of 'taking it that' as a weak form of belief… The psychologised explanation of the action is to be understood as the same explanation as the non-psychologised one. (Dancy 2000, 126)

Dancy makes two key observations: the first is that if the agent loses their belief that *p*, the possibility of explaining in terms of the fact that *p* disappears. The second is that the fact that *p* is not part of a (genuinely) different explanation of the agent's action from the fact that they believed that *p*.

The argument from impotent facts shows how Dancy's first observation, together with another general principle of explanation, entails his second. Applied to the case of my friend's award, it runs as follows: the fact that my friend had won an award would not have explained why I

congratulated her if I hadn't believed that she had won an award (because, *inter alia*, I would not have congratulated her). That being so, the explanatory power of the fact that my friend had won an award depends on my believing that she had. But one proposition cannot be part of a genuinely different explanation from another if the explanatory power of the former depends on the truth of the latter. Therefore:

> *Conclusion 1b* The fact that my friend won an award is not part of a genuinely different explanation of why I congratulated her from the fact that I believed that she had won an award.

The following sections set out this argument in more detail.

## 3.1 Impotence

If I hadn't believed that my friend had won an award I would not have congratulated her (I'm not a sarcastic sort[26]). So, in the event that my friend had won an award and I had not believed that she had, the fact that my friend had won an award would not explain why I had congratulated her. Thus:

> *Premise 1b* If, *ceteris paribus*, I had not believed that my friend had won an award (though she had[27]) then the fact that she had won an award would not have partially explained why I congratulated her.

This case is clear: had I not believed that my friend had won an award then the fact that she had won an award would not have explained why I congratulated her because *I wouldn't have congratulated her*. However, even if I had congratulated her, it would not have been because she won an award (absent some weird circumstances). For instance, suppose that I believed that she had just got a new job, I might have still congratulated her then, but even so it would be wrong to say that I congratulated her because she had won an award.

The point is that, in the absence of any weirdness[28], something that one would ordinarily take to make one's action worth doing could not explain why one did it *unless* one believed it. As others note:

---

[26] And, *ex hypothesi*, I didn't take anything else to make congratulating her, in some respect, worth doing.

[27] This remark is parenthetical because it is already implied by the *ceteris paribus* condition.

[28] Hornsby gives the following example of such weirdness: 'Consider George who is quite ignorant of the condition of the ice…It might be that George is sociable, and skates at the edge because that is where the other skaters are; and it might then be true that he skates at the edge because the ice in the middle is thin (there is a two-step explanation of Georges skating there which adduces the thinness of the ice).' (Hornsby 2007, 296) The point is that I'm not saying that a fact can *never* explain an agent's action unless the agent believes it. What I am establishing is that in either of the examples given (and we are kept 'in' those examples by the *ceteris paribus* clauses) – it does not explain my action unless I believe it.

> A fact cannot be a reason that explains one's action unless the person is aware of it. (Alvarez 2016a, 30)
>
> If I act in the light of the fact that I am married [i.e. because I am married], I must believe that I am. (Dancy 2000, 126)

So, the explanatory power of that which an agent believes typically depends upon their believing it.

## 3.2 The Difference Principle

When is one proposition part of a genuinely different explanation from another? Recall the case of overexplanation considered in § (VII)4.3: we said that the fact that swimming would improve my mood and the fact that swimming would help me sleep better were each parts of genuinely different explanations of why swimming was, in some respect, worth doing. In contrast, we said that the fact that Joanne's roof leaks and the fact that it rained last night were not parts of genuinely different explanations.

Why did we reach these conclusions? It was because even if it stopped being the case that swimming would help me sleep better, the fact that it would improve my mood would continue to explain why swimming was, in some respect, worth doing; and *vice versa*. In contrast, if Joanne's roof didn't leak, then, *ceteris paribus*, the fact that it rained last night would not explain why her carpet is wet. It was because of this difference between the two examples that we said the former involved genuinely different explanations, whereas the latter did not.

The point, I suggest, is this: a proposition is seemingly part of a genuinely different explanation from some other proposition only if its explanatory power does not depend on the truth of the latter. So the fact that it rained last night is not part of a genuinely different explanation of why the carpet is wet from the fact that the roof leaks because, *ceteris paribus*, if the roof didn't leak then the fact that it rained last night would stop explaining why the carpet was wet (it wouldn't be wet anymore).[29] Thus:

> DIFFERENCE For any propositions *p, q* and *r*, *p* is part of a genuinely different explanation of the fact that *r* from *q* only if, *ceteris paribus*, had *p* been the case and *q* not been the case, *p* would still partially explain the fact that *r*.

[29] Cf. 'If the rationalizing explanation is dependent on the physiological explanation in an appropriate sense (e.g., by being reducible to it), then in truth there is only one explanation here.' (Kim 1989, 80)

The difference principle connects the property of being part of a genuinely different explanation from some other proposition with the property of being logically independent of that other that proposition.[30]

## 3.3 Concluding the argument from impotent facts

Here, then, is what I have called 'the argument from impotent facts': the fact that my friend had won an award would not have explained why I congratulated her if I hadn't believed that she had. That being so, since some proposition is part of a genuinely different explanation from some other only if its explanatory power does not depend on the truth of the latter, the fact that my friend won an award is not part of a genuinely different explanation of why I congratulated her from the fact I believed that she had won an award.

Formally:

**The Argument from Impotent Facts**

| | |
|---|---|
| *Premise 1b* | If, *ceteris paribus*, I had not believed that my friend had won an award (though she had) then the fact that she had won an award would not have partially explained why I congratulated her. |
| DIFFERENCE | For any propositions *p, q* and *r*, *p* is part of a genuinely different explanation of the fact that *r* from *q* only if, *ceteris paribus*, had *p* been the case and *q* not been the case, *p* would still partially explain the fact that *r*. |
| *Conclusion 1b* | The fact that my friend won an award is not part of a genuinely different explanation of why I congratulated her from the fact that I believed that she had won an award. |

# 4 The argument for Premise 1

I said that Premise 1 can be inferred from the conclusion of the argument from false belief together with the conclusion of the argument from impotent facts, here's how: The argument from false belief tells us that there is a full explanation, call it 'Δ*', of why I congratulated my friend that includes the fact that I believed that she had won an award but not the fact that she had won an award. The argument from impotent facts tells us that the fact that my friend won an award is not part of a genuinely different explanation of why I congratulated her from the fact that I believed that she had won an award.[31]

Now, since Δ* is a full explanation of why I congratulated my friend that includes the fact that I believed that she had won an award, the argument from impotent facts means that the fact

[30] In the Appendix to § (X) I note that those whom I call 'inclusive disjunctivists' reject this principle.

[31] This means that, for any full explanation, Δ, of why I congratulated my friend that includes the fact that I believed that my friend had won an award, there is no full explanation that is both genuinely different from Δ and includes the fact that my friend won an award.

that my friend won an award cannot be part of an explanation of why I congratulated my friend that is genuinely different from Δ*. Therefore, there is a full explanation, Δ*, of why I congratulated my friend such that the fact that my friend won an award is neither a part of Δ* nor is it part of a full explanation that is genuinely different from Δ*.

Formally:

| | |
|---|---|
| *Conclusion 1a* | There is a full explanation of why I congratulated my friend that includes the fact that I believed that she had won an award but not the fact that she had won an award. |
| *Conclusion 1b* | The fact that my friend won an award is not part of a genuinely different explanation of why I congratulated her from the fact that I believed that she had won an award. |
| *Premise 1* | There is a full explanation of why I congratulated my friend such that the fact that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation. |

The formal argument for Premise 1 is set out in full in the Appendix to this chapter.

# 5 The Exclusion Principle

Premise 1 is the first premise of The Explanatory Exclusion Problem. The second premise is a final principle of explanation: *the exclusion principle*. The exclusion principle says that if some proposition is not part of a full explanation of some *explanandum*, and is not part of a genuinely different explanation of that *explanandum* then it does not explain that *explanandum*. In this section I set out the argument for the exclusion principle and, as an aside, discuss its relation to Kim's well-known *principle of causal exclusion*.

## 5.1 The argument for the exclusion principle

The reasoning behind the exclusion principle is straightforward: if you say that you can fully explain some *explanandum* without mentioning *p*, then *p* can't just be added to that full explanation (because it would be superfluous – and MINIMALITY precludes superfluous *explanans*), so *p* and that full explanation can't together be part of the same full explanation. Moreover, if *p* is also not part of a genuinely different full explanation, then we are drawn to the conclusion that, by the law of excluded middle, *p* is not a part of *any* full explanation of that *explanandum* (since it isn't part of the same full explanation and isn't part of a genuinely different explanation). But if it isn't part of any full explanation, then it isn't a partial explanation – which means, as I set out in the previous chapter, it does not explain.

Thus, if there is a full explanation of some *explanandum* that does include some fact then if that fact is not part of a genuinely different explanation of that *explanandum*, it does not explain it. Or, in other words:

> EXCLUSION For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.*

This is the second premise of The Explanatory Exclusion Problem.

## 5.2 The exclusion principle and the principle of causal exclusion

As an aside, before we conclude, it is worth noting that the exclusion principle is a close relation of a principle that is central to the exclusion problem in mental causation, namely, Kim's *principle of causal exclusion*:

> If an event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e* (unless this is a genuine case of causal overdetermination). (Kim 2008, 17)

There are several similarities between Kim's principle and mine.[32] However, despite my use of the 'exclusion principle' label, there are also some significant differences.

One difference that is particularly worth stressing is that while Kim's principle is restricted only to the consideration (or *exclusion*) of simultaneous events, there is no analogous restriction in my exclusion principle. This difference is particularly worth stressing because it is the reason why my argument against the exclusion principle (see § (XI)) does not also apply to Kim's principle of causal exclusion.

[32] In particular, much of the conceptual apparatus of Kim's principle of *causal* exclusion has an explanatory analogue in the conceptual apparatus I have used. For instance, Kim's (1993, 280) distinction between partial and sufficient causes, is, I suggest, the causal analogue of my distinction between partial and full explanation. It is thus possible to transpose Kim's principle into an explanatory analogue of it, using the structural principles of explanation I have assumed (it is perhaps worth noting here that Kim (1988, 233) originally formulated his principle as the principle of *explanatory* exclusion). Doing so reveals both the respects in which my exclusion principle is similar to his, and those in which it is not.

# 6 The Explanatory Exclusion Problem for (R1)

Here, then, is the argument of The Explanatory Exclusion Problem against the claim that I congratulated my friend because she won an award (i.e. against (R1)):

**The Explanatory Exclusion Problem for (R1)**

| | |
|---|---|
| *Premise 1* | There is a full explanation of why I congratulated my friend such that the fact that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation. |
| EXCLUSION | For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.* |
| *Conclusion 1* | The fact that my friend had won an award does not explain why I congratulated her. |

# 7 Conclusion

I have demonstrated how two seemingly trivial claims about what explains an agent's action when they act in error and in ignorance, together with five plausible principles of explanation can lead to the somewhat counterintuitive conclusion that I did not congratulate my friend because she had won an award, but only because I thought she did.

Indeed, as is presumably clear, this result should generalise beyond this example – The Explanatory Exclusion Problem is a problem for anyone who thinks that facts about things that are external to our minds can explain why we do the things that we do. So, for instance, if the Problem is to be believed then one never takes one's umbrella because it is raining, but only because one believes that it is raining. Similarly, one never waits to cross the road because a car is coming, but only because one thinks a car is coming.

The standard response to The Explanatory Exclusion Problem is to accept the conclusion and to insist that when I say that I congratulated my friend because she won an award, the *purported explanans* of that expression (the fact that my friend had won an award) is merely elliptical for the *real explanans*, which it conversationally implies (that I knew or believed that had she won an award). I will deal with this response, and other common responses in § (X). Before then, in the next chapter, I want to show how The Explanatory Exclusion Problem can be used to create further problems for explanatory rationalism.

# Appendix

## A.1 The argument for Premise 1

For reference, here is the argument for Premise 1, in full:

| | |
|---|---|
| *Premise 1a* | If, *ceteris paribus*, my friend had not won an award (but I still believed that she had), then (i) the fact that I believed that she had won an award would have partially explained why I congratulated her; and (ii) the fact that she had not won an award would not have partially explained why I congratulated her. |
| FACTIVITY | For any propositions *p* and *q*, if *p* partially explains the fact that *q* then *p* is the case. |
| ENDURANCE | For any propositions *p, q* and *r*, the following holds: Suppose that *q* partially explains the fact that *r* when it is not the case that *p.* Suppose further that neither *p* nor *not p* is part of the same explanation of *r* as *q*. Then, if, *ceteris paribus*, it were the case that *p*, *q* would still partially explain the fact that *r*. |
| SUFFICIENCY | For any proposition *q*, and any set, Δ, if Δ is a full explanation of the fact that *q* in some circumstance, then, in any circumstance in which all the elements of Δ partially explain the fact that *q*, Δ fully explains the fact that *q*. |
| *Premise 1b* | If, *ceteris paribus*, I had not believed that my friend had won an award (though she had) then the fact that she had won an award would not have partially explained why I congratulated her. |
| DIFFERENCE | For any propositions *p, q* and *r*, *p* is part of a genuinely different explanation of the fact that *r* from *q* only if, *ceteris paribus*, had *p* been the case and *q* not been the case, *p* would still partially explain the fact that *r*. |
| *Premise 1* | There is a full explanation of why I congratulated my friend such that the fact that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation. |

# (IX)

# Other Uses for The Explanatory Exclusion Problem

*In which I show how The Explanatory Exclusion Problem can be used to arrive at some other conclusions that are inconvenient for explanatory rationalism. I set out the general form of the Problem, followed by the general form of the argument for the first premise of the Problem. I show the Problem can be used to argue that the fact that I read that my friend had won an award does not explain why I congratulated her, and that neither that fact, nor the fact that she had won an award, can explain why it was* pro tanto *rational for me to congratulate her.*

Recall that in § (VI), I said that if explanatory rationalism is to be consistent with the *prima facie* reasonable claims set out in §§ (II)-(IV), the following must be true:

(R1) I congratulated my friend because she had won an award.

(R2) I congratulated my friend because I read that she had won an award

(R3) It was *pro tanto* rational for me to congratulate my friend because she had won an award.

(R4) It was *pro tanto* rational for me to congratulate my friend because I read that she had won an award.

In the previous chapter I showed how The Explanatory Exclusion Problem provides an argument against (R1). The purpose of this chapter is to show that The Explanatory Exclusion Problem also provides an argument against (R2), (R3) and (R4), by using it to reach the following conclusions:

*Conclusion 2* The fact that I read that my friend had won an award does not explain why I congratulated her.

*Conclusion 3* The fact that my friend won an award does not explain why it was *pro tanto* rational for me to congratulate her.

*Conclusion 4* The fact that I read that my friend had won an award does not explain why it was *pro tanto* rational for me to congratulate her.

# 1 The general form of The Explanatory Exclusion Problem

The general form of The Explanatory Exclusion Problem for the claim that some proposition, *x*, explains some proposition, *z*, is as follows:

**The general form of The Explanatory Exclusion Problem**

| | |
|---|---|
| *Premise #* | There is a full explanation of why *z* such that *x* is neither a part of that full explanation nor is it part of a genuinely different explanation. |
| EXCLUSION | For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.* |
| *Conclusion #* | x does not explain why z. |

The Explanatory Exclusion Problem can thus provide the argument for Conclusions 2, 3, and 4, if we provide the appropriate specification of 'Premise #'. How do we do so?

## 1.1 The general form of the argument for Premise #

In the argument for Premise 1 of the previous chapter, the only premises that were specific to the example considered were these:

| | |
|---|---|
| *Premise 1a* | If, *ceteris paribus*, my friend had not won an award (but I still believed that she had), then (i) the fact that I believed that she had won an award would have partially explained why I congratulated her; and (ii) the fact that she had not won an award would not have partially explained why I congratulated her. |
| *Premise 1b* | If, *ceteris paribus*, I had not believed that my friend had won an award (though she had) then the fact that she had won an award would not have partially explained why I congratulated her. |

The general form of these premises is, for the particular propositions *x, y* and *z,* as follows:

| | |
|---|---|
| *Premise #a* | If, *ceteris paribus*, *x* had not been the case (but *y* still had) then (i) the fact that *y* would have partially explained why *z*; and (ii) the fact that not *x* would not have partially explained why *z*. |
| *Premise #b* | If, *ceteris paribus*, *y* had not been the case (but *x* still had) then the fact that *x* would not have partially explained why *z*. |

All of the other premises in the argument for Premise 1 were, you will recall, general principles of explanation. As a result, if Premise #a and Premise #b are true of *x, y* and *z*, then Premise # is true of them too (given FACTIVITY, SUFFICIENCY, ENDURANCE and DIFFERENCE); and if Premise # is true of them then The Explanatory Exclusion Problem implies that *x* does not explain *z*. So, to arrive at Conclusions 2, 3, and 4 we need only show that the appropriate specifications of

Premise #a and Premise #b are true of those cases. I consider the argument for each conclusion in turn.

# 2 The Explanatory Exclusion Problem for (R2)

Does the fact that I read that my friend had won an award explain why I congratulated her? The Explanatory Exclusion Problem for (R2) concludes that it does not. To reach that conclusion we need to establish the following:

> *Premise 2* There is a full explanation of why I congratulated my friend such that the fact that I read that my friend had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation.

Now, as discussed, in order to arrive at Premise 2 we need only demonstrate that the appropriate specifications of Premise #a and Premise #b are true. That is the purpose of the following sections.

## 2.1 The argument for Premise 2a

Firstly, suppose that, *ceteris paribus*, I hadn't read that she had won an award, but I still believed that she had – maybe I saw her win it, or heard about it from another friend, or maybe (incredibly) I acquired the belief as the result of a brain aneurism. In such a circumstance, would I still have congratulated her? Of course I would! I thought that she'd an award! And, I submit, I would have congratulated her *because* I believed that she had won an award.

Moreover, continuing to suppose that, *ceteris paribus*, I hadn't read that she had won an award (but nonetheless believed that she had), would the fact that I *hadn't* read that she had won an award explain why I congratulated her? Surely not! Why would it? This is a prosaic case, not a weird one.

Thus, combining these two insights, we arrive at the following specification of Premise #a:

> *Premise 2a* If, *ceteris paribus*, I had not read that my friend had won an award (but I still believed that she had) then (i) the fact that I believed that she had won an award would have partially explained why I congratulated her; and (ii) the fact that I had not read that she had won an award would not have partially explained why I congratulated her.

## 2.2 The argument for Premise 2b

Now consider: if, *ceteris paribus*, I had not believed that she had won an award even though I'd read that she had in the newspaper (perhaps I'm sceptical of the mainstream media, or

jealousy makes me withhold), would I have congratulated her? Of course I wouldn't: as we've already established. I didn't think that she'd won an award, so it would have been odd of me to congratulate her (again, I didn't see anything else of worth in congratulating her, and I'm not a sarcastic sort).

But if I wouldn't have congratulated her then the fact that I had read that she had won an award wouldn't have explained why I congratulated her (since I wouldn't have, and *explananda* must be the case). Thus:

> *Premise 2b* If, *ceteris paribus*, I had not believed that my friend had won an award (although I had read that she had won an award) then the fact that I had read that she had won an award would not have partially explained why I congratulated her (since I wouldn't have).

## 2.3 The argument for Premise 2

To run through the argument, for clarity: we know, from condition (i) of Premise 2a, that if, *ceteris paribus*, I had not read that my friend had won an award (but had still believed that she had won an award) then there would have been a full explanation, call it Δ*, of why I congratulated her that would have included the fact that I believed that she had won an award. And, from FACTIVITY, we know that had I not read that she had won an award, the (false) proposition *that I read that she had won an award* could not have been a part of Δ*.

We also know, from condition (ii) of Premise 2a, that the fact that I *didn't* read that she had won an award would not have been a part of Δ*. So, since (had I not read that she had won an award) neither the (false) proposition *that I read that she had won an award* nor its (true) negation would have been elements of Δ*, we know, from ENDURANCE, that all the elements in Δ* must also have explained why I congratulated my friend in the case in which I *did* read that she had won an award. So, from SUFFICIENCY, we know that Δ* is likewise a full explanation of why I congratulated her when I *did* read that she had won an award.

Now, from Premise 2b, we know that if, *ceteris paribus*, I hadn't believed that my friend had won an award then the fact that I had read that she had won an award would not have explained why I congratulated her (since I wouldn't have). So, from DIFFERENCE, we know that the fact that I read that she had won an award cannot be part of a genuinely different explanation of why I congratulated her from the fact that I believed that she had won an award. And since Δ* includes the fact that I believed that she won an award, the fact that I read that she had won an award cannot be part of an explanation of why I congratulated her that is genuinely different from Δ*. Therefore:

*Premise 2* There is a full explanation of why I congratulated my friend such that the fact that I read that my friend had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation.

## 2.4 The Explanatory Exclusion Problem for (R2)

So, the Problem for (R2) is as follows:

**The Explanatory Exclusion Problem for (R2)**

*Premise 2* There is a full explanation of why I congratulated my friend such that the fact that I read that my friend had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation.

EXCLUSION For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.*

---

*Conclusion 2* The fact that I read that my friend had won an award does not explain why I congratulated her.

# 3 The Explanatory Exclusion Problem for (R3)

Does the fact that my friend won an award explain why it was *pro tanto* rational for me to congratulate her? The Explanatory Exclusion Problem for (R3) concludes that it does not. To reach that conclusion we need to establish the following:

*Premise 3* There is a full explanation of why it was *pro tanto* rational for me to congratulate my friend such that the fact that she won an award is neither a part of that full explanation nor is it part of a genuinely different explanation.

In the following sections I will demonstrate that the appropriate specifications of Premise #a and Premise #b are true.

## 3.1 The argument for Premise 3a

If, *ceteris paribus*, my friend had not won an award, but I still believed that she had, would it still have been *pro tanto* rational for me to congratulate her? I suggest that it would: in any normal circumstances if you think that your friend has won an award the rational thing to do is to congratulate her – if you didn't then you would be acting *irrationally* (unless, say, you were very jealous, or knew that she doesn't like to be congratulated – but that's not my example).

Why doesn't the falsity of my belief seem to matter? It is because, as Wedgwood notes:

> When we assess a choice or decision as rational or irrational, we are assessing it on the basis of its relation to the agent's beliefs, desires, and other such mental states – not on the basis of its relation to facts about the external world that could vary while those mental states remained unchanged. (Wedgwood 2002, 350)

For instance, we already acknowledged (see § (II)1.1) that it was rational for Sally to run given that she believed that a bear was chasing her[1]; furthermore, it was rational in spite of the fact that no bear was chasing her. What matters to the *pro tanto* rationality of an action is, as I've suggested, that the agent takes there to be something of worth in doing it – not that it actually is, in some respect, worth doing.

So, if, *ceteris paribus*, my friend had not won an award, but I still believed that she had, then it would have been *pro tanto* rational for me to congratulate her, and that would have been, in part, *because* I believed that she had won an award.

Moreover, in this counterfactual case, the fact that my friend had *not* won an award would clearly not have explained why it was *pro tanto* rational for me to congratulate her. Again, the case is, *ex hypothesi*, prosaic and not weird. Thus:

> *Premise 3a* If, *ceteris paribus*, my friend had not won an award (but I still believed that she had), then (i) the fact that I believed that she had won an award would have partially explained why it was *pro tanto* rational for me to congratulate her; and (ii) the fact that she did not win an award would not have partially explained why it was *pro tanto* rational for me to congratulate her.

## 3.2 The argument for Premise 3b

Now consider: if, *ceteris paribus*, I had not believed that she had won an award even though she had, would it still have been *pro tanto* rational for me to congratulate her? Perhaps you think that the answer depends on whether or not it was rational for me *not* to believe that she had won an award? I will return to the question of whether or not that matters in § 4.1, but suppose, for now, that it was.

If I didn't believe that she had won an award and didn't take congratulating her to be, in any other respect, worth doing, then it would not have been even *pro tanto* rational for me to congratulate her. No rational (non-sarcastic) person, with such beliefs and desires (etc.) would congratulate their friend. Thus:

> *Premise 3b* If, *ceteris paribus*, I had not believed that my friend had won an award (though she had) then the fact that she had won an award would not have partially explained why it was *pro tanto* rational for me to congratulate her (since it wouldn't have been *pro tanto* rational for me to congratulate her).

[1] If you are concerned about whether or not her belief is rational in the first place, please forestall those concerns until § 4.1.

## 3.3 The Explanatory Exclusion Problem for (R3)

Trusting that the reasoning is now familiar, I will spare the reader a demonstration of how I think we can arrive at Premise 3 from Premise 3a and Premise 3b. So, the Problem for (R3) is as follows:

**The Explanatory Exclusion Problem for (R3)**

| | |
|---|---|
| *Premise 3* | There is a full explanation of why it was *pro tanto* rational for me to congratulate my friend such that the fact that she won an award is neither a part of that full explanation nor is it part of a genuinely different explanation. |
| EXCLUSION | For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.* |
| *Conclusion 3* | The fact that my friend won an award does not explain why it was *pro tanto* rational for me to congratulate her. |

# 4 The Explanatory Exclusion Problem for (R4)

Finally: does the fact that I read that my friend had won an award in the newspaper explain why it was *pro tanto* rational for me to congratulate her? The Explanatory Exclusion Problem for (R4) says it does not. This is what we need to show to get there:

| | |
|---|---|
| *Premise 4* | There is a full explanation of why it was *pro tanto* rational for me to congratulate my friend such that the fact that I read that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation. |

## 4.1 The argument for Premise 4a

Suppose, as we did in § 2.1, that *ceteris paribus*, I hadn't read that my friend had won an award, but I still believed that she'd won an award. If I hadn't read that she'd won an award, but, say, I'd seen her win it, would it still have been *pro tanto* rational for me to congratulate her? Of course it would. Likewise, if I'd heard about her award from a (reliable) friend it would have been *pro tanto* rational for me to congratulate her. But what if I acquired the belief as the result of brain aneurism? Would it still have been *pro tanto* rational for me to congratulate her? Some think it would not, for instance:

> If an agent has irrational beliefs, those beliefs are not able to make rational any actions done in their light. (Dancy 2000, 60)

> It would not be rational for Holly to put on winter clothes if her belief that it is snowing were due to crazed conviction, say, or wishful thinking. Irrationality cannot beget rationality! A subject's beliefs contribute to making it rational for her to act in certain ways only if those beliefs are themselves rational. (Whiting 2014, 4)

If this is so, then it seems that the way that one acquires one's beliefs *is* relevant to the rationality of one's actions; that is, we cannot so easily omit mention of them in the explanation of why it is rational for an agent to do some action.

In some respect, I disagree – I think that an irrational belief can nonetheless explain why it was *pro tanto* rational (but probably not *all things considered* rational) for someone to do something,[2] as do others.[3] Nonetheless, even if irrational beliefs *can't* explain why actions are *pro tanto* rational, that does not mean that the experiences or appearances on which an agent's beliefs are based are necessary to a full explanation of why it is *pro tanto* rational for that agent to do some action; it means only that the fact that the relevant beliefs of the agent are rational[4] is necessary to a full explanation.

So, if *ceteris paribus*, I hadn't read that my friend had won an award, but I still believed that she'd won an award *and* that belief was (still) rational, would it have been *pro tanto* rational for me to congratulate my friend? Of course! And it would have been *pro tanto* rational, in part, because I believed that she had won an award.

Moreover, the fact that I did *not* read that my friend had won an award clearly would not explain why it would have been *pro tanto* rational for me to congratulate her. Thus:

---

[2] An argument to that effect: suppose that Bernard Ortcutt, spy extraordinaire, comes to believe that the FBI has discovered that he is a spy and has sent agents to apprehend him. His belief is well-founded – FBI counter-intelligence exists to capture spies like him; a normally regular asset has gone missing; and he's been repeatedly trailed by a black sedan this week. It is, I submit, *pro tanto* rational for Bernard to go into hiding (since he is a patriot (so isn't minded to turn), and has no interest in jail time, it is likely also *all things considered* rational). Now consider Ornard Bertcutt: Ornard has actually lived a rather pedestrian life but, through some freak co-incidence (a peculiar mental disorder, say), his mental states are all identical to Bernard's – he is Bernard's mental duplicate. So, Ornard, like Bernard, believes that the FBI is out to get him. However, Ornard's belief is not only false, it is plainly not rational. Nonetheless, it is, I submit, at least *pro tanto* rational for him to go into hiding. Some resistance to this view is understandable: to believe that the FBI is chasing you is certainly outlandish, and Ornard only believes it because he is crazy – so surely going into hiding can't be even a *pro tanto* rational thing for him to do? If Ornard goes into hiding it's because he's *not* rational, it's not the rational thing for him to do! The problem, however, with saying that it isn't even *pro tanto* rational for Ornard to go into hiding is that we are seemingly forced to question whether or not it is *pro tanto* rational for Bernard to go into hiding. Why? Because of the widely held view that what it is rational for an agent to do supervenes on their mental states (e.g. Broome 2013, 151), which is to say that there can be no change in what it is rational for an agent to do without a change in their mind. Now, since Bernard and Ornard are mental duplicates, given that what it is rational for an agent to do supervenes on their mental states, there can be no difference between Bernard and Ornard in what it is rational for them to do, since there is no difference in their brain states.

[3] 'Given my irrational belief that smoking will protect my health, it would be rational for me to smoke. Given this hermit's irrational belief that his life of self-inflicted pain would please God, he could rationally live such a life.' (Parfit 2011, 114)

[4] Whatever your preferred standard of rationality for beliefs is.

*Premise 4a* If, *ceteris paribus*, I had not read that my friend had won an award (but I still believed that she had won an award) then (i) the fact that I believed that she had won an award would have partially explained why it was *pro tanto* rational for me to congratulate her; and (ii) the fact that I had not read that she had won an award would not have partially explained why it was *pro tanto* rational for me to congratulate her.

## 4.2 The argument for Premise 4b

Finally: if, *ceteris paribus*, I had not believed that she had won an award even though I'd read that she had, would it have been *pro tanto* rational for me to congratulate her? Again, perhaps you think it depends on whether or not it was rational for me not to believe that she had won an award. And perhaps, further, you insist that given the *ceteris paribus* clause it can't have been rational. Well, I disagree.

Supposing that pathological jealousy makes me withhold. The *all things considered* rational thing for me to do is to cease withholding (and perhaps seek treatment). Then, *once I've done that*, it would be *pro tanto* rational for me to congratulate my friend. But, given that I don't believe that she has won an award (that is, before I cease withholding), I suggest, it can't be *pro tanto* rational for me to congratulate her. I see nothing of any worth in doing so – and not because I don't like her, or don't care about her feelings – but because I *don't* believe that she has won an award. To insist that it is even *pro tanto* rational for me to congratulate her, despite the fact that I don't believe that she has won an award, is to insist that it is *pro tanto* rational for me to do something that I take to be, *in no respect*, worth doing. I don't see how that could be rational.

So, given that it wouldn't have been *pro tanto* rational for me to congratulate her, the fact that I read that she had won an award could not explain why it would have been *pro tanto* rational for me to congratulate her. Thus:

*Premise 4b* If, *ceteris paribus*, I had not believed that my friend had won an award (although I had read that she had won an award) then the fact that I had read that she had won an award would not have partially explained why I it was *pro tanto* rational for me to congratulate her (since it wouldn't have been *pro tanto* rational for me to congratulate her).

## 4.3 The Explanatory Exclusion Problem for (R4)

Again, I will not spell out the reasoning from Premise 4a and Premise 4b to Premise 4. So, concluding, the Problem for (R4) is:

**The Explanatory Exclusion Problem for (R4)**

| | |
|---|---|
| *Premise 4* | There is a full explanation of why it was *pro tanto* rational for me to congratulate my friend such that the fact that I read that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation. |
| EXCLUSION | For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.* |
| *Conclusion 4* | The fact that I read that my friend had won an award does not explain why it was *pro tanto* rational for me to congratulate her. |

# 5 The Argument from Illusion

Many have observed[5] a similarity between the problem that I have formalised as The Explanatory Exclusion Problem and the argument from illusion in the literature on perception.

To the extent that we conceive of the argument from illusion as a problem for the idea that the external world could explain why we believe what we believe then the argument from illusion can be represented as an instance of The Explanatory Exclusion Problem, briefly, as follows.

I see a tomato so I believe that there is a tomato before me. If no tomato had been before me, but it still appeared to me as though one had, then the fact that it appeared to me as though a tomato were there would partially explain why I believed that a tomato was there. And the fact that there wasn't a tomato wouldn't. This provides the relevant specification of Premise #a.

And then if there had been a tomato but it hadn't appeared to me as though there had (blind spots in my vision, say), then I wouldn't have believed that there was a tomato, so the fact that there was a tomato wouldn't have explained why I believed that there was (since I wouldn't have believed that there was). This provides us with the relevant specification of Premise #b.

From these specifications of Premise #a and Premise #b, together with the relevant principles of explanation, we can, in a manner that should now be familiar, arrive at The Explanatory Exclusion Problem, and the consequent conclusion that the fact that there is a tomato before does not explain why I believe that there is.

However, there are other versions of the argument from illusion according to which it is not a claim about what *explains* an agent's beliefs. For instance, one interpretation of it is as a problem for the idea that veridical perceptual experience provides a basis for knowledge.

[5] (E.g. Stout 1996; Dancy 2000; Hornsby 2008; Hyman 2011; McDowell 2013)

When the argument from illusion is conceived in that manner, it cannot be so simply characterised as The Explanatory Exclusion Problem. Although, I submit, the problems are nonetheless related.

# 6 Conclusion

I have now shown how The Explanatory Exclusion Problem can provide arguments against (R1)-(R4).

Since explanatory rationalism requires the truth of (R1)-(R4) if it is to be consistent with the *prima facie* reasonable claims set out in earlier chapters, I will need to find some way to reject the conclusions of The Explanatory Exclusion Problem. My solution will be to reject the exclusion principle, which, I will argue, by means of several counterexamples, is clearly false. I will argue, *inter alia*, that the fact that my friend won an award explains why I congratulated her because it explains why I believed that she won an award, and the exclusion principle fails to provide for the transitivity of that sort of explanation.

Before then, however, I wish to consider the other, more commonplace responses to The Explanatory Exclusion Problem, and what is wrong with them. That is the focus of the next chapter.

# (X)

# How normative reasons don't explain

*In which I reject two accounts of how normative reasons explain. I re-introduce talk of normative reasons, defining them as things that make actions, in some respect, worth doing. I ask how it is that we manage to explain our actions when we say that we acted because of a normative reason there was to act; for instance: how is it that I explain why I took my umbrella when I say that I took it because it was raining? I suggest that the fact that it was raining explains why I took my umbrella either 'elliptically', 'directly' or 'indirectly'. I note that which answer one accepts will depend on one's response to The Explanatory Exclusion Problem: elliptical theorists accept the conclusion of the Problem, direct theorists reject the first premise, and indirect theorists reject the second. I set out the problems with elliptical and direct theories.*

We often explain why we do something by citing some fact that counted in favour of doing it: I say that I took my umbrella because it was raining; Marshall says that he is going to the station because his daughter is on the 7 o'clock train. How are we to make sense of these commonplace explanations?

If The Explanatory Exclusion Problem is to be believed, we should not interpret them literally. That is, the *purported explanantia* (i.e. that which follows the 'because') in such statements are not the *actual explanantia*; the fact that it is raining does not *really* explain why I took my umbrella. Instead, whatever explanatory power these statements have is due to there being a short-hand; when I say that I took my umbrella because it was raining, the fact that it was raining is merely *elliptical* for what really does the explaining, which is the fact that I believed that it was raining, or that I knew that it was.

Alternatively, one could reject the conclusion of The Explanatory Exclusion Problem, and insist that these remarks are literally accurate; the fact that it is raining really does explain why I took my umbrella. There are two ways of doing this: either by rejecting the first premise of the Problem, or the second.

Theories that reject the first premise say that the contribution of the fact that it is raining to the explanation of why I took my umbrella is, in some sense, *independent from*, or *in addition to* the explanatory contribution of the fact that I believed that it was raining. They argue that the fact that it is raining explains my action *directly* (that is, unmediated by features of my psychology). According to the most popular theories, the direct explanatory relation between

the world and the action is the result of the special connection between the world and actions that knowledge engenders.

In contrast, theories that reject the second premise accept that the explanatory contribution of the fact that it was raining is not *in addition to* the explanation that is already provided by, *inter alia*, the fact that I believe that it was raining. Thus they deny that the fact that it is raining *directly* explains why I took my umbrella. However, they insist, the fact that it was raining does *indirectly* explain why I took my umbrella, by explaining why I believed that it was raining.

In short: we want to understand how it is that I manage to explain my action when I say, 'I took my umbrella because it was raining.' There are three possible accounts: the fact that it was raining either explains my action *elliptically*, or it explains it *directly*, or it explains it *indirectly*.

In what follows I will set out the problems with *elliptical* and *direct* theories, which are the typical responses to this problem. In subsequent chapters I will defend my own *indirect* theory. Before then it will help me better characterise what is at stake in this discussion if we re-introduce talk of 'normative reasons'; that is the focus of the first section.

# 1 Normative reason explanations

## 1.1 Normative reasons

I want to re-habilitate the term 'normative reason', which I abandoned in § (I). Let us define it as follows:

> *Definition* For any *p*, *p* is a normative reason for *A* to *φ* if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing.[1]

A few points, already discussed in § (I)4, are worth stressing here: Firstly, while this definition is closely aligned to the conventional definition of normative reasons, it departs from it in so far as I am not saying anything about how normative reasons, so defined, relate to expressions like 'the reasons there are to act' or 'the reasons for which an agent acted' – for my purposes the term 'normative reason' is strictly a term of art meaning anything that makes an action, in some respect, worth doing.

Secondly, I will assume, to avoid ambiguity, that 'counting in favour of' and 'making, in some respect, worth doing' are equivalent relations. This assumption has no bearing on the argument of this chapter.

[1] Note: this definition makes no explicit assumptions about the ontology of '*p*'.

Finally, it is worth recalling that that which counts in favour of an action (i.e. that which makes it, in some respect, worth doing), is typically *not* a feature of the agent's psychology, which is to say that normative reasons are typically not features of an agent's psychology.

## 1.2 Normative reason explanations

In § (VII)1, I noted that the word 'explains' and 'explanation' have different meanings. If we say that the fact that it rained last night is an *explanation* of why Joanne's carpet is wet, we are giving one sense of 'explanation', in which an explanation is an *explanans*. This is the sense of 'explanation' and 'explains' that I have focussed on in previous chapters – it is the sense that is involved in the concepts of full and partial explanation.

However, I suggest that 'I took my umbrella because it was raining' is an *explanation* of why I took my umbrella in different sense of the word – the sentence is an *elucidation* of why I took my umbrella, it is not the *explanans* of why I took it. It is in that sense of the word 'explanation' that I suggest that we call sentences in which a normative reason for an agent to act appears in the position of an *explanans* of why they acted,[2] 'normative reason explanations'.

It is a fact that we often give normative reason explanations of our actions. I gave a few examples in my opening remarks, here are some more[3]: Sandra is going to the shops because she is out of milk; I'm flying to Bodrum because that's where my father lives; Theresa May made a deal with the DUP in 2017 because that was the only way for her to form a majority government.[4] In all of these examples, something that made the action, in some respect, worth doing (for that agent), which is to say, a normative reason for them to do it, appears to explain their action. And, indeed, when such explanations are given, you understand why the action was done – so an explanation of some sort has certainly been provided.

# 2 Theories of normative reason explanation

In the introduction to this chapter I asked how it was that I managed to explain my why I took my umbrella when I said that I took my umbrella because it was raining. The remarks of the previous section should have made clear that this is a specific instance of a more general

[2] That is, sentences like '*A φ'*d because *p*' or 'the fact that *p* explains why *A φ'*d' where the fact that *p* is a normative reason for *A* to *φ*.

[3] For the following examples, assume that the relevant supporting conditions (desires, evaluative judgements, evaluative facts…) are in place such that: the fact that Sandra is out of milk is a normative reason for her to go to the shops; the fact that my father lives in Bodrum is a normative reason for me to go there; the fact that making a deal with the DUP was the only way for May to form a majority government was a normative reason for her to do so.

[4] Of course, the *purported explanantia* of these remarks are, at best, partial explanations.

question: how do we manage to explain our actions when we give a normative reason explanation of them? Answering this question is the job of what I will call a 'theory of normative reason explanation'.

Which theory of normative reason explanation one holds depends on one's response to The Explanatory Exclusion Problem: those who accept the conclusion of the Problem insist that when we give a normative reason explanation it is not really the normative reason that explains our action. Instead, they argue, the normative reason is elliptical for that which really does the explaining – which is the agent's awareness of, or belief in the normative reason. This is the *elliptical* theory of normative reason explanation.

Those who reject the first premise of the Problem insist that, when we give a normative reason explanation of an agent's action, the normative reason explains the agent's action *directly*; that is, the explanatory relations involved are unmediated by features of an agent's psychology. These are *direct* theories of normative reason explanation.

Finally, those who reject the second premise of the Problem accept the primacy of features of the agent's psychology in explaining their action, but nonetheless insist that normative reasons can explain an agent's action *indirectly*, by explaining those features of the agent's psychology that explain their action. These are *indirect* theories of normative reason explanation.

The focus of the next two sections is on critiquing *elliptical* and *direct* theories of normative reason explanation, respectively. Subsequent chapters are devoted to the defence of my own *indirect* theory of normative reason explanation.

# 3 Elliptical theories

Elliptical theories[5] accept the conclusion of The Explanatory Exclusion Problem; they suggest that when I say that I took my umbrella because it was raining, the *purported explanans* (the fact that it was raining) is not the *actual explanans*. Nonetheless, when I say that I took my umbrella because it was raining, you understand why I took it – that is, in spite of the apparent inaccuracy of what I said, you still understood why I took my umbrella. The question is: how? How do we manage to explain our actions if the *purported explanans* in a normative reason explanation is not the *actual explanans*?

The elliptical theorist's response, is to say that, in a normative reason explanation, the normative reason is elliptical for the *actual explanans*, so, for instance, when I say that I took my umbrella because it was raining, the fact that it was raining is elliptical for some feature of

[5] The name for these theories was inspired by Maria Alvarez's (2010, 180) related (but not identical) discussion of what she calls 'Humean explanations.'

my psychology, which is what really does the explaining. That is, when we give a normative reason explanation, 'we suppose that, properly understood, it should be seen as enthymematic, i.e. as an acceptable shorthand version of the full explanation.' (Dancy 2000, 121)

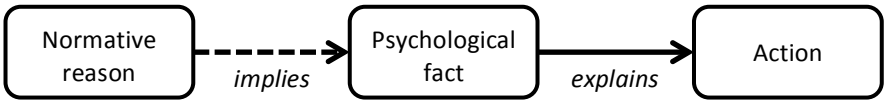


**Figure X-1:** Elliptical theories of normative reason explanation

Although they are rarely[6] explicitly advocated, elliptical theories provide what is probably the *de facto* account of how facts about the world explain our actions.

## 3.1 Elliptical for what?

What feature of our psychology is it that normative reasons are meant to be elliptical for, when we give a normative reason explanation? Opinions diverge. One view is that if I say that I took my umbrella because it was raining the conversational implicature[7] is the fact that I believed that it was raining,[8] another view is that it is the fact that I acted for the reason that it was raining, [9] and another still is that it is the fact that I knew that it was raining.[10] Of these three views only the latter is robust to a particular sort of challenge posed by Gettier cases,[11] so I will assume that the conversational implicature of saying that I took my umbrella because it was raining is that I knew that it is raining.

[6] Sandis (2013) and, I think, Dancy (2014) are rare exceptions. Sandis (2012, 178 fn. 24) also attributes this view to Michael Smith.
[7] Which, recall, is what really explains why I took my umbrella.
[8] This is probably the *de facto* view, and, I think, is explicitly the position of Sandis (2013).
[9] See Dancy (2014).
[10] Gibbons (2010, 359) comes close to advocating this view – although I think his eventual position is closer to the *indirect* theory that I advocate in § (XII), since, on the same page, he makes clear that normative reasons can explain.
[11] The challenge is this: recall, from § (IV)1.4 that when Edmund had a Gettier belief (i.e. justified and true but not knowledgeable) that the ice was thin we could not say that he stayed at the edge *because* the ice was thin. However, if the conversational implicature of saying 'He stayed at the edge because the ice was thin' is merely that he believed that the ice was thin, then there is no reason why we should not say it – this is then a problem for elliptical theories that take the *purported explanans* to be elliptical for the fact that the agent believed it. In contrast, if the conversational implicature is that he *knew* the ice was thin, then we should not say that he stayed at the edge because the ice was thin, since he did not know it. Dancy's (2014) account, according to which the conversational implicature is that he stayed at the edge *for the reason that the ice was thin* could account for the fact that we don't say that he stayed at the edge because the ice was thin if he were willing to say that an agent only acts for the reason that *p* if they know it; but he isn't, so it can't.

## 3.2 The problems with elliptical theories

Elliptical theories claim that normative reason explanations provide some explanation of an agent's action only because they imply that the agent knew that normative reason. Critically, elliptical theorists insist, the normative reason does no explanatory work of its own.

I raise two related concerns with this view: first, it makes normative reasons explanatorily inert, contrary to the prevailing view that they ought to have explanatory power; and second, it renders ordinary language explanations of our actions thoroughly unsuited to the task to which we habitually put them.

On the first: the 'explanatory constraint', so named by Jonathan Dancy (2000),[12] and to which many[13] subscribe, says that any theory of reasons must account for someone doing some action *because* of a normative reason that there is for them to do it. This constraint seems like a modest one: assuming that normative reasons indeed have some normative import, the explanatory constraint requires only that normative reasons have more than *just* normative significance. As Ulrike Heuer succinctly puts it, this requirement, 'expresses nothing more than the everyday assumption that we sometimes… do something because it is right or justified.' (Heuer 2004, 47) Indeed, one must wonder what the point of normativity is if we can't do things *because* there are such normative reasons for us to do them (why recommend an action if that recommendation can't affect whether or not you do it?).

The problem for elliptical theories is that they clearly fail the explanatory constraint. They hold that it is never the normative reason *per se* that explains the agent's action, but only what is implied by it – thereby rendering normative reasons explanatorily inert.[14]

The second problem: denying that normative reasons explain our actions contradicts our habitual patterns of speech. We routinely cite normative reasons by way of explanation of our actions and it does quite severe disservice to our ordinary language expressions to suppose that when I say, 'Laura threw away the milk because it had gone off' the real *explanans* is not what I say it is[15] but only what is implied by it. Ordinary language may be occasionally imprecise or misleading, but to accept such ubiquitous misrepresentation as a part of our everyday accounts of actions seems to be a high price to pay. I think we can do better.

---

[12] Dancy's work is, to my knowledge, also the first appearance of this argument.

[13] (e.g. Dancy 2000, 101; Smith 2004, 175; Hornsby 2007, 301; Raz 2009, 194; Hieronymi 2011, 415)

[14] This argument is similar to The Right Reasons Problem (see § (III)3.2). However, while that argument required that an agent should be able to do something *for reasons that* make it worth doing, this argument requires that an agent should be able to do something *because of* what makes it worth doing, indeed, they should be able to do it because *it is worth doing*. The problem for The Explanatory Exclusion Problem is that it is incompatible with the idea that such facts could explain an agent's action.

[15] That is, what follows the 'because'.

# 4 Direct theories

The only way to satisfy the explanatory constraint and to accept the literal form of our everyday expressions is to concede that (non-psychological) normative reasons can explain our actions. This means rejecting the conclusion of The Explanatory Exclusion Problem. The first way to do that, which we consider now, is to reject the first premise.

Recall the general form of the first premise of The Explanatory Exclusion Problem:

> *Premise #* There is a full explanation of why *z* such that *x* is neither a part of that full explanation nor is it part of a genuinely different explanation.

Now, as we established in § (VIII), it seems as though whenever we give a normative reason explanation we can make a claim that fits the Premise # form about the explanatory contribution of the normative reason to the explanation of the action[16]; that is, for any action and any normative reason to do that action, there is always a full explanation of that action such that that normative reason is neither a part of that full explanation nor is it part of a genuinely different full explanation. Direct theories deny this claim.

## 4.1 How normative reasons *directly* explain actions

Direct theories say that when a normative reason explains an action it adds something to the explanation of that action that is independent of, and in addition to, what the fact that the agent knew it provides.

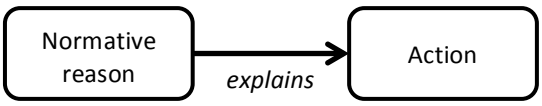


**Figure X-2:** Direct theories of normative reason explanation

But how are facts about the world supposed to *directly* explain our action? Here is a typical response: following Gilbert Ryle (1949), one can conceive of knowledge as a capacity or ability. In particular, one can conceive of knowledge as the capacity to respond to a fact about the world, that is, the capacity to respond to a normative reason.

The suggestion is that, when an agent acts from knowledge:

- 'The fact that things actually are the way they believe them to be weighs with them' (Hornsby 2008, 254); or
- The fact exerts a rational influence on the agent's will (McDowell 2013); or

[16] There are, perhaps, exceptions in 'weird' cases, where the fact that *p* and the fact that the agent believes that *p* are both normative reasons for them to *φ* (see Dancy 2000, 124) – but these are peculiar enough that they can be ignored.

- The agent is guided by the fact (Hyman 2015); or
- The agent exhibits a rational response to the fact. (Smith 2004)

Thus, knowing some fact engenders a special, direct connection between the agent and the fact. When one knows something the fact itself guides one or impresses itself upon one's action and thereby accounts for what one does: that is how normative reasons *directly* explain an agent's action.

Not all direct theories rely on knowledge to account for the direct connection, however. Some, such as Dancy's (2000) non-factive theory of normative reason explanation do not think that a normative reason need even be true in order to explain an agent's action, so long as it was believed. More generally, even amongst knowledge-based direct theories, the precise nature of *how* the normative reason explains depends on the way in which one rejects the first premise of The Explanatory Exclusion Problem. I consider the main ways of being a *direct* theorist in the Appendix to this chapter.

## 4.2 The problems with direct theories

### 4.2.1 What's weird about *direct* normative reason explanation

How exactly does 'responding' to a normative reason, being 'guided' by one or 'acting in light of it' make that normative reason *directly* explanatory? The nature of the explanation that knowledge is supposed to engender is thoroughly mysterious, and accounts of it are replete with metaphors but thin on detail. If the concept of 'responding to the fact' is not causal (and none seem to think it is), what exactly is the nature of the direct connection between the agent and the fact, when they respond to the fact, that makes that fact explain their action?[17]

Now, if it were self-evident that there is such a direct connection then the use of metaphor might well be unproblematic. For instance, we can say that maglev trains are 'guided' by magnets without needing to be literal about the relation between the magnets and the train because it is seemingly clear that something the magnets are doing is directly affecting the train. However, it is very much not clear that normative reasons are *directly* affecting my action – so we need an account of what is occurring that is not couched in metaphors in order to convince us that the normative reason really is directly related to the action. I am concerned that no such account is available because there is no such direct relation. This is the first problem for direct theories.

[17] Similar remarks can be made for direct theories that aren't knowledge-based, such as non-factive theories. See remarks in the Appendix to this chapter for further detail.

### 4.2.2 You cannot credibly reject Premise 1

Now for the second problem for direct theories. In §§ (VIII)2-4, I set out the argument for the following claim:

> *Premise 1* There is a full explanation of why I congratulated my friend such that the fact that she had won an award is neither a part of that full explanation nor is it part of a genuinely different explanation.

Direct theories reject this claim, in this instance, and they reject that a similar sort of claim can be made for any normative reason that purports to explain an agent's action. The problem, I suggest, is that there is no credible way to reject Premise 1.

First, recall that Premise 1 followed from four principles of explanation together with the following two claims:

> *Premise 1a* If, *ceteris paribus*, my friend had not won an award (but I still believed that she had), then (i) the fact that I believed that she had won an award would have partially explained why I congratulated her; and (ii) the fact that she had *not* won an award would not have partially explained why I congratulated her.
>
> *Premise 1b* If, *ceteris paribus*, I had not believed that my friend had won an award (though she had) then the fact that she had won an award would not have partially explained why I congratulated her.

Assuming that these claims are uncontentious,[18] the only options for one who wants to reject Premise 1 are to reject one of FACTIVITY, ENDURANCE, SUFFICIENCY, or DIFFERENCE.

I think that there is no unproblematic way for a direct theorist to reject one of these principles of explanation. I provide a full account of my reasoning in the Appendix to this chapter, however, by way of overview here: denying FACTIVITY comes at the cost of denying something that is seemingly obviously true (i.e. that explanation is factive). Denying DIFFERENCE makes the concept of being a genuinely different explanation obscure, and relying on the denial of difference to account for normative reason explanation results in an implausibly ubiquitous level of overexplanation. Meanwhile, theories that deny ENDURANCE seemingly stretch credulity by insisting that, in some cases, the mind plays no explanatory role in action. And lastly, theories that deny SUFFICIENCY must insist that normative reasons are indispensable to the explanation of an action if they are known – and there is no good account of why that should be the case.

[18] And I know of no one who would deny either.

The upshot, I suggest, is that there is no credible way to reject the first premise of The Explanatory Exclusion Problem for normative reason explanation, and therefore no credible way of being a direct theorist.

# 5 Conclusion

I have discussed two possible responses to The Explanatory Exclusion Problem for normative reason explanation, elliptical theories and direct theories. I suggested that both of these theories are deeply problematic.

If normative reasons don't explain *elliptically,* and don't explain *directly*, then, I suggest, they must explain *indirectly*, that is, by explaining those features of the agent's psychology that explain their actions. In the next chapter I will provide the basis for my indirect theory by showing that the exclusion principle is false. Subsequent chapters will then set out and defend my indirect theory.

# Appendix

## A.1 Four direct theories of normative reason explanation

Direct theories of normative reason explanation reject the idea that the first premise of The Explanatory Exclusion Problem is always true of normative reasons. As noted in § 4.2.2, this means that direct theories must reject one of the following principles of explanation:

FACTIVITY — For any propositions *p* and *q*, if *p* partially explains the fact that *q* then *p* is the case.

DIFFERENCE — For any propositions *p, q* and *r*, *p* is part of a genuinely different explanation of the fact that *r* from *q* only if, *ceteris paribus*, had *p* been the case and *q* not been the case, *p* would still partially explain the fact that *r*.

ENDURANCE — For any propositions *p, q* and *r*, the following holds: Suppose that *q* partially explains the fact that *r* when it is not the case that *p.* Suppose further that neither *p* nor *not p* is part of the same explanation of *r* as *q*. Then, if, *ceteris paribus*, it were the case that *p*, q *would* still partially explain the fact that r.

SUFFICIENCY — For any proposition *q*, and any set, Δ, if Δ is a full explanation of the fact that q *in* some circumstance, then, in any circumstance in which all the elements of Δ partially explain the fact that q, Δ fully explains the fact that *q*.

Since any direct theory must reject one of these principles, we can categorise the different direct theories according to which of these principles they reject, as follows[19]:

- *Non-factivist*: Theories that reject FACTIVITY and insist that when normative reasons explain an agent's action they do so *qua* the content of the agent's belief.

- *Inclusive disjunctivist*: Theories that reject DIFFERENCE and insist that when normative reasons explain what an agent does they do so as part of an explanation that is genuinely different from the explanation in terms of the agent's psychology.

- *Exclusive disjunctivist*: Theories that reject ENDURANCE and insist that when normative reasons explain what an agent does they do so as part of the full explanation *instead of* the facts about what an agent believes.

- *Supplementarist*: Theories that reject SUFFICIENCY and insist that when a normative reason explains an agent's action they do so as part of the full explanation *together with* the facts about what an agent believes.

In what follows I will set out the account of how normative reasons explain in each of these strategies, followed by the problems they face.

A caveat regarding all these theories: they are all to a greater or lesser extent, of my own construction. Most of the literature from which these theories are drawn is actually a discussion of whether or not (and how) normative reasons can be *the reason for which an agent acts*. I, however, am answering a simpler question: can normative reasons explain our actions? The theories below have been inspired by responses to the former question, and while I suggest that certain authors hold some forms of the theories below, nothing in what I say depends on these attributions being accurate. That is, so long as they are correct characterisations of possible theories, it is not vital that those to whom I attribute them would agree with the attribution.

## A.2 Non-factivist theories

Non-factivist theories reject the factivity principle. According to non-factivist theories (e.g. Dancy 2000; Comesaña and McGrath 2014), saying that I took my umbrella because it was raining has the conversational implicature that it was raining, but does not entail that it was raining; indeed, they stress, that implicature is cancellable.

[19] There is a fourth strategy that I haven't considered here. One could differentiate the *explanantia* of normative reasons and psychological facts (I believe that Stout (1996) adopts this strategy, but one could also develop such a strategy based on Hornsby (2008) and McDowell's (2013) disjunctivism about acting) – and consequently argue that normative reasons can explain *something* an agent does e.g. the fact that there is beer in the fridge explains *the fact that I got beer* (I couldn't have done so had there not been beer there). I don't consider this strategy here because I think it suffers all the failings of elliptical theory without its simplicity – we want to be able to give a normative reason explanation of the fact that I went to the fridge, not merely the fact that I got beer.

These theories suggest that when a normative reason explains an agent's action facts about what the agent believes are really second fiddle in explanatory terms to the normative reason: they act as mere *enabling conditions* for the normative reason (*qua* what the agent believes) to explain what the agent does.

The idea seems to be this: a normative reason cannot explain an agent's action unless the agent believes it[20] – so the fact that the agent believes it must at least *enable* the explanation. However, 'there is a difference between a consideration that is a proper part of an explanation, and a consideration that is required for the explanation to go through, but which is not itself a part of that explanation.' (Dancy 2000, 127)

According to this strategy, the facts about what the agent believes don't have the explanatory force necessary to explain the agent's action, only the normative reason does. That is, the fact that I believe that it is raining enables the contents of my belief, namely, *that it is raining*, to explain why I took my umbrella (and so they are independent parts of the same explanation – although it's the normative reason that is, in some sense, the major party). They say that we should understand expressions like 'Sally ran because she thought a bear was chasing her' *appositionally*, where the reference to Sally's believing is a qualification on the truth value of the *explanans*, but not a part of the *explanans* itself.[21]

Non-factivist theories can thus accept that the same full explanation is available whether the agent's belief is true or false *because* the truth or falsity of the content of the agent's belief has no bearing on whether or not it is in that full explanation.

### A.2.1 The problem for non-factivist theories

The first problem with non-factivist theories is that they are absurd. If we admit that falsehoods can explain that seems to make the concept of explanation itself thoroughly mysterious. If someone says that they took their umbrella because it was raining even though it wasn't, I would not think they were cancelling a Gricean implicature, I would think that they misspoke, because what they said is just plainly contradictory.

Secondly, non-factivist theories rest their claim to plausibility on the idea that beliefs are mere enabling conditions – they don't have the requisite explanatory force in and of themselves. I am sceptical that the perceived explanatory weakness of 'enabling conditions' is genuine and isn't merely a consequence of their salience in a given context. Were it merely a question of salience then the explanatory power of facts about the agent's psychology would be restored

[20] Recall the discussion of § (VIII)3.1.
[21] Dancy proposes this sort of reading: 'Sally ran because a bear was chasing her, as she believed.'

and we would again have to question what normative reasons add by way of explanation. But even putting such scepticism aside, it is, as Turri (2009) notes, far from clear that facts about what an agent believes are aptly categorised as enabling conditions.[22] And if they aren't enabling conditions, concerns about what normative reasons really do by way of explanation return.

Thirdly, non-factivist theories struggle to explain why it is that, when an agent has a Gettier belief, we don't give a normative reason explanation of their action.[23] For instance, we don't say that Edmund stayed at the edge because the ice was thin when his justified belief that it was thin is only accidentally true – although the non-factivist would insist that we should.[24]

Fourthly, it is worth noting, from a rhetorical perspective, that even the progenitor of non-factivist theories, Jonathan Dancy, has since abandoned them (see Dancy 2014).

## A.3 Inclusive disjunctivist theories

Inclusive disjunctivist theories reject the difference principle. They accept that there is a full explanation of why I congratulated my friend that does not include the fact that she won an award.[25] Nonetheless, they say, there is a genuinely different explanation of why I congratulated my friend that *does* include the fact that she won an award. They suggest that even though the explanation in terms of normative reasons would not exist without the explanation in terms of beliefs, it is nonetheless genuinely different from it.[26]

What would make you believe this? Well, as already noted,[27] there are structural similarities between the argument from illusion in the literature on perception and the problem for the

[22] Turri (2009, 505–6) argues that it is normally odd to ask why an enabling condition for an explanation obtains, but that it is not normally odd to ask why an agent believed what they believed – therefore facts about what an agent believes aren't enabling conditions.
[23] See fn. 11.
[24] See, for instance, Dancy (2014, 89) for this criticism of non-factivist theories.
[25] This is what makes them *inclusive disjunctivists* as opposed to the *exclusive disjunctivists*, to be considered in the next section. I have adopted the inclusive vs. exclusive distinction from Ruben (2008) and Stout (2009); they say that inclusive disjunctivists accept some role for the highest common factor, while exclusive disjunctivists do not. Note Pautz (2010, 298–99) draws the same distinction between different types of disjunctivism, but he calls inclusive disjunctivism the 'overdetermination version'; and exclusive disjunctivism and 'the restrictive version'.
[26] This is, I think, the view that John Hyman sets out: 'If James merely believed that going to church would please his mother but did not know that it would, we can say that he went to church because he believed that it would please his mother, but we cannot say that he went to church because it would please his mother. But if he knew that it would please his mother, we can say either that he went to church because he knew that it would please his mother or that he went to church because it would please his mother.' (Hyman 2011, 366–67)
[27] See § (IX)5.

explanation of action that I have construed as The Explanatory Exclusion Problem.[28] One could, inspired by disjunctivist responses to the former, adopt a disjunctivist view of action explanation as a response to the latter. The idea is this: although it is always explanatory, the highest common factor of all action explanations (explanation in terms of the facts about what the agent believes) is not the limit of the resources available for action explanation[29] – sometimes we can *also* explain an agent's action directly, and *differently*, with a normative reason. So while the action is already fully explained by facts about what the agent believes (and relevant supplementary facts (e.g. facts about what they want, judge good etc.)) the normative reason adds another, additional explanation of it.

### A.3.1 The problems for inclusive disjunctivist theories

The first problem with inclusive disjunctivist theories is that, in rejecting the difference principle we are rejecting a seemingly plausible account of (at least part of) what it is that makes explanations genuinely different. As a result, I find, we begin to lose our handle on what it is that makes the explanation genuinely different.

Secondly, and relatedly, it's not at all clear what would be in the full explanation of the agent's action that includes the normative reason. Given MINIMALITY, it cannot include all the elements of the full explanation that includes the belief – so it must be some other set of facts. But what other set of facts is also sufficient for me to congratulate my friend and which explains my doing so *directly*? I cannot think of any.

And lastly, and perhaps most obviously, the inclusive disjunctivist accepts that my action is *overexplained*, because there are two genuinely different explanations of it. This is problematic because, as Dancy puts it, '[it] would mean that there are somehow too many explanations around' (2000, 171).[30] And this is not a benign case of overexplanation – it's not at all clear that the fact that I congratulate my friend is explained in two different *ways*.

[28] Consider: For perception, when it, for instance, appears as though there is a tomato one is always in either one of two states: either one is seeing a tomato or it *merely* appears to one as though there is a tomato. The fact that it appears as though there is a tomato is the highest common factor of these two states. Compare Horsnby's (2008) construal of action: when one acts on the belief that it is raining one is always in one of two states – acting on the knowledge that it is raining or *merely* acting on the belief that it is raining. Acting on the belief that it is raining is seemingly the highest common factor of these two states.

[29] As McDowell puts it: 'The point of the disjunctive approach is to reject a highest common factor conception, not in the sense of denying that there are common features between the disjuncts, but in the sense of refusing to restrict our resources for rational explanation to those that are available for the "worse" disjunct.' (McDowell 2013, 27)

[30] Davis (2005) also makes this argument.

The problem isn't just confined to this case either. What the inclusive disjunctivist requires is that *whenever* we can give a normative reason explanation of an action, the agent's action is overexplained. But if this sort of overexplanation is meant to occur whenever we give a normative reason explanation, then, given the ubiquity of such explanations, inclusive disjunctivism seems to require a really implausible incidence of overexplanation.

Maybe you want to insist that overexplanation should not trouble us as much as overdetermination, because there aren't *independent determining* factors, there are only *different explanatory* factors. However, I think this response is more trouble than it's worth – it's not clear to me that it makes overexplanation unproblematic and, as I've noted, we start to lose our grasp on what makes genuinely different explanations *genuinely different* if they aren't, in some sense, independent explanations.

## A.4 Exclusive disjunctivist theories

Exclusive disjunctivist theories reject the endurance principle. They reject the idea that, when I congratulate my friend, the fact that I believed that she had won an award explains why I congratulated her. According to this stronger sort of disjunctivism (hence *exclusive* rather than *inclusive*) either an agent acts because they believe the world to be a certain way or they act because the world *is* that way, but never both. On this account, when a normative reason explains an agent's action it does so *instead of* the facts about what an agent believes.[31]

Why would you believe this? That is, why think that sometimes beliefs don't explain? Stoutland makes the following argument for this view:

> If someone goes to a room because a meeting is being held there, that is adequate justification for her effort. If it turns out that the meeting isn't there, we have to revise our justification, and hence our explanation, and say she went to the room because she *believed* the meeting was there. Her belief becomes an explanatory factor, that is to say, just in case she was mistaken about the situation originally appealed to as justification. This is the general case: beliefs become explanatory factors when agents are mistaken about the situations originally taken to justify them.
>
> The situation is analogous to someone's flipping a switch to turn on a light, without the light going on. In this case he *tried* to turn on the light, just because he failed, which would not be

[31] Both Collins (1997) and Stoutland (1998) adopt this response to the Problem. I may have mis-interpreted Hyman's (2011) view in fn. 26, in which case I think this is his position. Sandis (2012, 119) also attributes this view to Alvarez (2010), though I'm not convinced it is her position. In particular, Alvarez agrees that beliefs still explain in veridical cases (cf. 'It is always *possible* (and sometimes necessary, namely when the agent acted on a false belief) to give explanations in the psychological form.' (Alvarez 2013, 149)). More generally, it's not clear to me that Alvarez is responding to The Explanatory Exclusion Problem as I have characterised it – her discussion focuses on the pragmatic considerations that determine whether or not one gives a normative reason explanation or one in terms of psychological facts. However, to the extent that she does respond to the redundancy objection, I think she is better characterised as adopting the supplementarist theory.

> true if he succeeded in turning it on without difficulty. From the fact that what someone did when he *failed* to turn on a light is that he *tried*, it doesn't follow that what he did when he succeeded also included trying. Analogously, given that what explains my going to a room where I do *not* find my friend is that I *believed* my friend was there, it does not follow that what explains my going to the room when my friend *is* there is also that I *believed* she was there. (Stoutland 1998, 61)

I think that the most coherent reading of Stoutland's account is as a rejection of the idea that whenever an agent acts, they act on the way they take things to be. That is: either an agent acts on the way they take things to be *or*, to use McDowell's (2013) phraseology, they 'act in light of the facts', but not both.[32] If they act on the way they take things to be, then the fact that they take them to be that way explains their action, but if they act in light of the facts (i.e. in light of a normative reason), then it is the facts that explain their action. That is, either facts about what they believe explain their actions *or* normative reasons explain their actions, but not both.

### A.4.1 The problems for exclusive disjunctivist theories

Firstly, by rejecting the endurance principle the exclusive disjunctivist is committed to the view that when some irrelevant false proposition becomes true, that can destroy pre-existing partial explanation relations. For the reasons set out in (VIII)2.3.1, that seems implausible.

However, even if it turns out that the endurance principle is false, the exclusive disjunctivist theory is a difficult position to maintain because there is widespread support for the view that beliefs always play some role in explaining what an agent does.

Exclusive disjunctivism insists that there is no highest common factor that is relevant to the explanation of action, whether one is right or mistaken. That is, it insists that there is no significant common factor between someone who acts on something they know to be the case and someone who acts on something they *merely* believe. I find this incredible.

[32] This reading has Stoutland endorsing (exclusive) disjunctivism about *acting*. This is different from disjunctivism about what explains an agent's action, or what their reason for acting is. It's worth noting that Hornsby (2008) and McDowell (2013) give a considerably more thorough account of disjunctivism about acting, but theirs is of an *inclusive* kind – they concede that whenever an agent acts they act on the way they take things to be. So while I take Stoutland to be a disjunctivist about the same thing, I don't invoke them here since his disjunctivism is far stauncher. There is also an alternative reading (perhaps truer to his precise statements) on which Stoutland embraces an exclusive disjunctivism about *action explanation* (i.e. an exclusive version of the disjunctivism considered in the previous section). All that such a disjunctivism insists is that the facts about what an agent believes do not always explain their action, it is silent on the question of whether or not an agent always acts on the way they take things to be. This latter sort of disjunctivism is weaker than the former in so far as it is entailed by it. I find the latter sort less coherent without the former since it is not clear to me how one could act on the way one takes things to be without the fact that one takes them to be that way playing some role in explaining what one does; for that reason I don't consider it further.

Consider: when I believe that it's raining I am either in a situation in which I know that it's raining or I *merely* believe (i.e. without knowing) that it is raining. Given that I take my umbrella whichever of the two situations I am in, doesn't the possibility that the two situations may be indistinguishable to me just mean that I act on the way I take things to be in both situations?[33] And given that I act on the way I take things to be, the fact that I take them to be that way must explain my action. I find positions that deny this reasoning impossible to believe.

Moreover, Stoutland's supposed argument by analogy is not so much an argument for his view as it is merely the application of it to another area. It seems clear to me (and to others (e.g. O'Shaughnessy 1973), that someone who goes to turn on a light *tries* to turn on the light whether or not they end up doing so.[34] Stoutland's rejection of that view is just the same sort of (exclusive) disjunctivism as his rejection of the view that one always acts on the way one takes things to be.[35] And one who is not persuaded of his view in the latter is unlikely to be persuaded by the application of it somewhere else.

Of course, my incredulous stare may do nothing to alter the opinion of someone who believes such a theory; however, I don't think I am alone in my incredulity. Surely there are theories that are easier to believe?

## A.5 Supplementarist theories

Finally, supplementarist strategies reject the sufficiency principle. They argue that just because facts about what an agent believes play a role in explaining their action that does not mean that normative reasons are not needed for a full explanation of the agent's action; indeed, they argue, normative reasons are an independent (and necessary) part of the *same* explanation as the facts about what an agent believes.[36]

Supplementarist strategies differ from non-factive strategies in that they insist that the normative reason can explain the agent's action only if it is true; for instance, the supplementarist might insist, the normative reason only explains an agent's action if the agent knows it. What the supplementarist rejects is the idea that the same full explanation of an

[33] This is not a denial of disjunctivism *tout court* – for instance, an inclusive disjunctivist (e.g. McDowell 2013) can acknowledge that one always acts on the way one takes things to be; they just deny that acting on the way one takes things to is *all* that a person ever does.
[34] Compare these commonplace expressions: 'he tried and failed'; 'he tried and succeeded' – Stoutland's account makes the former true of anyone who tries, and makes the latter a contradiction.
[35] Indeed, Dancy (2008b) discusses a disjunctivist account of trying to act.
[36] As noted in fn. 31 to the extent that Alvarez (2010) gives a response to The Explanatory Exclusion Problem, I think this may be her view.

agent's action is available whether or not the agent knows what they believe. Instead, they say, the normative reason is an indispensable part of the explanation when (and only when) it is known. Importantly, the supplementarist does not violate MINIMALITY because they do not think that the set that omits the normative reason is still sufficient to explain the agent's action when the agent acts from knowledge of the normative reason.[37]

### A.5.1 The problems for supplementarist theories

The first problem for supplementarist theories is that in rejecting the sufficiency principle they reject a seemingly plausible principle of explanation. That is, if some set of partial explanations is sufficient to explain some *explanandum* in one case, given that whenever those partial explanations all explain an *explanandum* they are sufficient to entail it, why would they also not be sufficient to explain it?

However, even if there were a good argument for rejecting the sufficiency principle, it is not clear why it should be false *in this case*. That is: it's not clear why knowing a normative reason makes that normative reason an indispensable part of the full explanation of the agent's action. Given that *what I believe* doesn't explain my action when I am mistaken, so an explanation in terms of psychological facts suffices, why is that explanation insufficient (and the normative reason indispensable) when my belief happens to knowledgeable?

I suppose the supplementarist theorist will answer this question by saying that it's because, when the agent knows the normative reason, the normative reason is, itself, a part of the story of why they acted because *it guides them* or *they respond to it*. When an agent doesn't know the normative reason, it doesn't guide them, so it isn't part of the reason why they acted. Then I fall back to my first concern (see § 4.2.1): I just don't know what it means for a normative reason to guide someone and until I do, I can't see why I should accept that normative reasons are indispensable to an explanation only if they are known.

[37] Although they do violate C-MINIMALITY (see § (VIII) fn. 23).

# (XI)

# The Exclusion Principle is False

*In which I show that the exclusion principle is false. I provide two counterexamples to the exclusion principle, one involving causal explanation and another involving non-causal explanation. I suggest that they are counterexamples because in each case the purportedly excluded fact explains the* explanandum *by explaining something that, in turn, explains the* explanandum*. I suggest that the problem with the exclusion principle is that it discriminates against all but the most proximal explanations of any given* explanandum*, and that this is problematic at least partly because we are typically interested in more distal explanations. I explain where our reasoning went wrong and which full explanation an apparently excluded fact is part of.*

In the previous chapter I argued that we should not accept the conclusion of The Explanatory Exclusion Problem (at least as far as normative reason explanations are concerned). I also argued that we should not reject the first premise of the Problem. If we should not accept the conclusion and we should not reject the first premise, the only remaining response to the Problem is to reject the second premise, the exclusion principle. The purpose of this chapter is to show that the exclusion principle is false and to explain why it is false. The next chapter builds on these insights to provide the makings of an indirect theory of normative reason explanation.

## 1 Two counterexamples to the exclusion principle

Recall the exclusion principle:

EXCLUSION For any propositions, *p* and *q*, if there is a full explanation of why *q* such that *p* is neither a part of that full explanation nor is it part of a genuinely different explanation, then *p* does not partially explain *q.*

In this section I want to discuss two counterexamples to EXCLUSION, one involving causal explanation and the other non-causal explanation.

### 1.1 A causal counterexample

Jean contracts HIV after having been transfused with infected blood. It goes undiagnosed for so long that, tragically, he develops AIDS. This much seems clear: Jean developed AIDS, in part, because he was transfused with HIV-infected blood. That is:

(a) The fact that Jean was transfused with HIV-infected blood partially explains why he developed AIDS.

However, we can also give the following explanation of why he developed AIDS: he had HIV and it went untreated. The fact that he had HIV and the fact that it went untreated are, I suggest, parts of a single full explanation of why he developed AIDS. Is the fact that he contracted HIV from an infected blood transfusion a part of that explanation? I suggest not: the fact that he had HIV is already enough[1] for Jean to develop AIDS,[2] so the fact that he was transfused with HIV-infected blood would be superfluous in any full explanation that already included the fact that he had HIV.[3] Thus, there is a full explanation of why he developed AIDS that includes the fact that he had HIV but does not include the fact that he was transfused with HIV-infected blood.[4]

Moreover, the fact that Jean was transfused with HIV-infected blood only explains why he developed AIDS *given* that he had HIV.[5] So, from DIFFERENCE, the fact that he was transfused with HIV-infected blood cannot be part of a genuinely different explanation of why he developed AIDS from the fact that he had HIV.

So, it follows that:

> (b) There is a full explanation of why Jean developed AIDS such that the fact that he was transfused with HIV-infected blood is neither a part of that full explanation nor is it part of a genuinely different explanation.

Now notice that (b) is the antecedent condition of EXCLUSION while (a) is the denial of the consequent condition. That is, the fact that Jean was transfused with HIV-infected blood explains why he developed AIDS in spite of the fact that there is a full explanation of why he developed AIDS that it is neither a part of and nor is part of a genuinely different explanation. So, given that (a) and (b) are true, EXCLUSION must be false.

---

[1] Together with the fact that it went untreated and the fact that a sustained, untreated HIV infection leads to AIDS, etc.

[2] We can see this by noting that a full explanation of this sort (he had HIV, it went untreated, untreated infections lead to AIDS…) would have been available, *ceteris paribus*, however he contracted HIV.

[3] And MINIMALITY precludes the possibility of full explanations having superfluous elements.

[4] We can arrive at this conclusion using the established reasoning from previous chapters, as follows: if, *ceteris paribus*, Jean had not been transfused with HIV-infected blood but still had HIV (never mind how he contracted it), then he would still have developed AIDS, and the fact that he had HIV would have (partially) explained why he developed AIDS. Moreover, in this counterfactual circumstance, the fact that he had *not* been transfused with HIV-infected blood would not have been part of the explanation of why he developed AIDS. This supplies us with the appropriate specification of Premise #a, which, together with FACTIVITY, ENDURANCE and SUFFICIENCY, leads us to the conclusion that there is a full explanation of why he developed AIDS that includes the fact that he had HIV but does not include the fact that he was transfused with HIV infected blood.

[5] If, *ceteris paribus*, Jean had not contracted HIV *despite* having been transfused with HIV-infected blood then he wouldn't have developed AIDS, so the fact that he was transfused with HIV-infected blood would not have explained *why* he developed AIDS. This supplies us with the appropriate specification of Premise #b.

## 1.2 A non-causal counterexample

Perhaps you agree that EXCLUSION is false for causal explanation, but you insist that it is true of non-causal explanation. If that were true, and one maintained that reason explanation is non-causal, then The Explanatory Exclusion Problem for normative reason explanation would return. The purpose of this section is to demonstrate that EXCLUSION is equally false for non-causal explanation.

Consider the following chess opening:

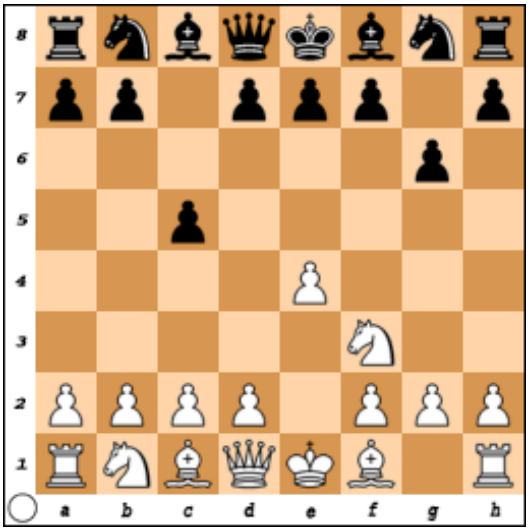

**Figure XI-1:** The Hyper Accelerated Dragon

This game has just begun. White's king and kingside rook are in their starting positions (e1 and h1, respectively). It is typical in the early stages of a chess game for a player to seek to castle. For a player playing white, castling on her king's side always involves moving the king from e1 to g1, and (as part of the same turn) the rook from h1 to f1. However, White cannot castle on her king's side because of the bishop (on f1) between the king and the kingside rook, which obstructs the move. That is:

> (c) The fact that there is a bishop between the king and the kingside rook partially explains why White cannot castle on her king's side.

I take it that the explanatory relationship here is clearly non-casual (or at least not causal in the familiar sense in which Jean's story is a causal story).

Now, the World Chess Federation's handbook states that:

> Castling is prevented temporarily… if there is any piece between the king and the rook with which castling is to be effected. (FIDE, n.d., 3.8.2.2)

This fact (concerning the rules of chess), together with the fact that there is a *piece* between White's king and her kingside rook fully explains why White cannot castle on her king's side.

But that full explanation does not include the fact that there is a *bishop* between the king and the kingside rook.[6] So, it seems, there is a full explanation of why White cannot castle on her king's side that includes the fact that there is a *piece* between the king and the rook but does not include the fact that there is a *bishop* between her king and king's side rook.[7]

Moreover, the fact that there is a bishop between the king and the rook only explains why White cannot castle if there is a *piece* between the king and the rook.[8] That being so we know (from DIFFERENCE) that the fact that there is a bishop between the king and the rook is not part of a genuinely different explanation of why White cannot castle from the fact that there is a piece between the king and the rook.

So, it follows that:

> (d) There is a full explanation of why White cannot castle such that the fact that there is a piece between the king and the rook is neither a part of that full explanation nor is it part of a genuinely different explanation.

Now notice, again, that (d) is the antecedent condition of EXCLUSION while (c) is the denial of the consequent condition. So, given that (c) and (d) are true, and given that the explanatory relations involved are non-causal, EXCLUSION must be false of even non-causal explanation.

## 2 Why these are counterexamples to the exclusion principle

Why are these cases counterexamples to the exclusion principle? Answering that question means answering the question of how the seemingly excluded facts nonetheless explain the *explananda* they are excluded from. My answer is straightforward: they explain by explaining those facts that *in turn* explain the *explananda*.

---

[6] Again, the latter is just superfluous to that full explanation, and MINIMALITY precludes such superfluity.

[7] Again, we could argue to this conclusion using the conventional reasoning: If, *ceteris paribus*, there had been no bishop between the rook and the king but there had still been a *piece* between the king and the rook (suppose that there was a knight there instead), then the fact that there was a piece between the king and the rook would still have explained why White could not castle. Moreover, had there been no bishop between the king and the rook, the fact that there was no bishop between the king and the rook would not have explained why White could not castle. This provides us with the relevant specification of Premise #a, so, from FACTIVITY, ENDURANCE and SUFFICIENCY we can conclude that there is a full explanation of why White cannot castle that includes the fact that there is a piece between the king and the rook but does not include the fact that there is a bishop between the king and the rook.

[8] If the bishop hadn't been a piece (perhaps one is playing a chess variant that excludes bishops, but for which the rules are otherwise the same), then White would have been able to castle, so the fact that there is a bishop between the king and the rook would not have explained why White *wouldn't* have been able to castle (since White would have been able to castle). This is the relevant specification of Premise #b. Consider: if there had been a penny on the board between the rook and the king we would not say that White cannot castle; however, in the chess variant in which bishops aren't pieces, they are, from the game's perspective, no different from pennies – that is, they are just irrelevant.

## 2.1 Revisiting the causal counterexample

In the first counterexample I insisted that the fact that Jean was transfused with HIV-infected blood explains why he developed AIDS. How does it do so? It does so by explaining why he has HIV.

Of course, the fact that Jean has HIV is the *more immediate* explanation of the fact that he developed AIDS. And, of course, the way one contracts HIV makes no difference to whether or not one develops AIDS *given that one has HIV*. If your question were why, *given that he had HIV*, did Jean develop AIDS, then the relevant explanation includes the fact that it went untreated, and the facts about how a sustained, untreated HIV infection leads to AIDS – and that explanation need make no mention of how he contracted HIV. But it would be odd to always take the fact that he had HIV as given! And once you don't take it as a given, a part of the explanation of how he came to develop AIDS is *why he has HIV in the first place*. And he has HIV because he was transfused with HIV-infected blood.

Summarising: the fact that Jean was transfused with HIV-infected blood explains both the fact that he has HIV and the fact that he developed AIDS, and the fact that he has HIV also explains the fact that he developed AIDS. Symbolically: *p* partially explains *q*, *q* partially explains *r* and *p* partially explains *r*. The question is: why does *p* partially explain *r*? I think the simplest and most natural answer is that the explanatory relations involved are transitive.

Given the transitivity of the (partial) explanation relations involved, the fact that *p* explains *q* together with the fact that *q* explains *r* ensures that *p* explains *r*. So, the fact that Jean was transfused with HIV-infected blood explains why he developed AIDS because it explains that which explains why he developed AIDS and because of the transitivity of the explanatory relations involved.

## 2.2 Revisiting the non-causal counterexample

What of the chess example? What is the relationship between the fact that there is a piece between the king and the kingside rook and the fact that there is a bishop between the king and the kingside rook?

Well, the fact that there is a piece between the king and the kingside rook is an existential fact. And existential facts are explained by their instances:

> 'Why is it that something is *F*? Because *A* is *F*. An existential quantification is explained by providing an instance.' (Lewis 1987, 223)

> 'Existential quantifications are true because of their true instances.' (Schnieder 2011, 460)

So, why is there a piece between the king and the kingside rook? Because there is a bishop between the king and the kingside rook (and a bishop is a piece…) – the fact that there is a bishop between the king and the kingside rook explains why there is a piece between the king and the kingside rook.

Summarising: the fact that there is a bishop between the king and the kingside rook explains *both* why there is a piece between the king and the kingside rook *and* why White cannot castle on her king's side, and the fact that there is a piece between the king and the kingside rook *also* explains why White cannot castle. Symbolically, again: *p* partially explains *q*, *q* partially explains *r* and *p* partially explains *r*. And, again, I think the most natural explanation of these facts is that the explanatory relations involved are transitive. That is: *p* explains *r because p* explains *q* and *q* explains *r*.

# 3 What's wrong with the exclusion principle

Having understood how the apparently excluded facts of these two counterexamples nonetheless explain, we are now in position to see what was wrong with the exclusion principle.

First, some basic terminology: if *p* explains *q* and *q* explains *r* and the explanatory relations involved are transitive then we can say that *p* is a *distal* explanation of *r*, while *q* is a more *proximal* explanation of *r*. So, the suggestion of the previous section was this: the fact that Jean was transfused with HIV-infected blood and the fact that there is a bishop between White's king and her kingside rook are *distal* (partial) explanations of their respective *explananda*. Whereas, by comparison, the fact that Jean had HIV and the fact that there is a piece between the king and the kingside rook are more *proximal* (partial) explanations of their respective *explananda*. I suggest that the problem with the exclusion principle, and the reason why it is wrong, is that it denies the explanatory power of *distal* explanations.

It will always be true of *any* distal (partial) explanation of some *explanandum* that there is a full explanation of that *explanandum* of which it is not part. Moreover, a distal explanation does not explain the *explanandum* in a way that is *genuinely different* from the more proximal explanation. So it will seemingly always be true of any distal explanation that there is a full explanation of some *explanandum* such that that distal explanation is not a part of that full explanation nor is it part of a genuinely different full explanation. That is, a distal explanation of some *explanandum* will always satisfy the antecedent condition of EXCLUSION.

That being so, the only way to preserve EXCLUSION is to deny the explanatory status of *distal* explanations – but that it is a very heavy price to pay. It means insisting that only the *most* proximate explanation of some *explanandum* explains it. And this is absurd!

Suppose we say that Franz developed lung cancer because he smoked. This explanation is, in some sense, an *extremely* distal explanation of why he developed lung cancer. A more proximal explanation would be that he regularly inhaled carcinogens. A more proximal explanation still would be the fact that cells in his lungs mutated.[9]

What each of these more proximal explanations have in common is that a full explanation of the fact that Franz developed lung cancer can be given in terms of them which makes no mention of, nor even entails, the more distal explanation. For instance, given that Franz was inhaling that mix of carcinogens, it doesn't matter (*ceteris paribus*) whether he got them from smoking or from passive smoking. Likewise, given that the cells in his lungs mutated it doesn't matter (*ceteris paribus*) to his developing lung cancer whether the cell mutation was the result of carcinogen inhalation or exposure to radiation. So, if the exclusion principle is to be believed, we can't even say that Franz developed cancer because he regularly inhaled carcinogens. According to the exclusion principle, only the *most* proximal explanation of his lung cancer explains it. This is surely absurd.[10]

The absurdity of the exclusion principle should be clear when we see that even an agent's beliefs can be excluded from the explanation of their actions. Suppose that, when I congratulate my friend, I believe that congratulating her will make her happy and, say, I desire to make her happy (or what have you). Given that I believe that congratulating her will make her happy, the explanation of why I believe it is excluded from the explanation of why I

[9] 'The formation of covalent bonds between the carcinogens and DNA producing DNA adducts, and the resulting permanent mutations in critical genes of somatic cells is the major established pathway of cancer causation by cigarette smoke.' (Hecht 2006, 609)

[10] Yablo (2008) makes an analogous criticism of Russel's remarks about causation, which follow:

> If the cause is a process involving change within itself, we shall require… causal relations between its earlier and later parts; moreover it would seem that only the later parts can be relevant to the effect… Thus we shall be led to diminish the duration of the cause without limit, and however much we may diminish it, there will still remain an earlier part, which might be altered without altering the effect, so that the true cause… will not have been reached. (Russell 1917, 135)

Russell suggests that only the 'later part' (i.e. the more proximal part) of the cause can be taken to be the cause itself, since the 'earlier part' (i.e. the more distal part) of the cause can obtain without guaranteeing the effect, if the later part does not obtain. This is the causal analogue of (what I take to be) the implication of the exclusion principle. In response to Russell's thesis about what it takes to be a cause, Yablo notes that, 'if this…were truly disqualifying…essentially everything would be robbed of its intuitive causal powers.' (2008, 298) My point is that the exclusion principle has exactly the same implication for explanation: if it were true then essentially everything would be robbed of its intuitive *explanatory* power.

congratulate her. That is, I don't need to believe that she has won an award to believe that congratulating her will make her happy, and if I don't believe that it would make her happy I wouldn't congratulate her, even if I believed that she had won an award. The fact that I believe that congratulating her would make her happy is a *more proximal* explanation of why I congratulate her. But it is absurd to infer from this that the fact that I believe that she had won an award does not explain why I congratulated her!

The exclusion principle is a damaging explanatory prejudice – discriminating against all but the most proximal explanations of any given *explanandum* will actually impede our ability to offer the explanations that we ordinarily give, because we are typically interested in more distal explanations of any given *explanandum* than the most immediately proximal explanation. For instance, if you're looking to prevent lung cancer then it matters that Franz developed lung cancer *because* he smoked – knowing that cell mutation explains his cancer doesn't help you much. What's most wrong with the exclusion principle, then, is that it forces us to say that the explanations we are interested in aren't really explanations.[11]

# 4 Where did we go wrong?

Supposing you accept my reasoning, some questions still remain: if the fact that Jean was transfused with HIV-infected blood explains why he developed AIDS then it must be part of some full explanation of why he developed AIDS. However, as established in (b), there is a full explanation of why he developed AIDS such that the fact that he was transfused with HIV-infected blood is neither a part of that explanation nor is it part of a genuinely different explanation. So how could it be part of any explanation? It was exactly this line of reasoning that, in § (VIII)5, led us to the exclusion principle. Since we have established that the exclusion principle is false, we should see what was wrong with this reasoning.

The mistake, I suggest, was the implicit assumption that if two full explanations of some *explanandum* are not identical then they must be genuinely different explanations of that *explanandum*. The point to recognise is that for two full explanations to be genuinely different requires more than just non-identity, it requires, *per* the difference principle, some form of independence. That is, the answer to the question of how it could be part of *any* explanation is that two full explanations can be non-identical without being genuinely different. For instance, there are two full explanations of the fact that Jean developed AIDS such that, although they

[11] Recall that in § (VIII)5.2, I noted that my argument against the exclusion principle is not an argument against Kim's principle of causal exclusion (in spite of the apparent similarity of their names and forms). This is because, as may quite be clear, Kim's principle does not exclude distal causation since it is only restricted to the exclusion of simultaneous causes.

are not *genuinely different* explanations of why he developed AIDS, they are nonetheless not identical; and, in particular, the fact that he was transfused with HIV-infected blood is an element of one and not the other.

# 5 Which explanation are distal explanations part of?

Another question: which full explanation of why he developed AIDS includes the fact that Jean was transfused with infected blood? We know it cannot be the same full explanation as the one that includes the fact that he had HIV, because how he contracted HIV is redundant in that explanation. So which full explanation is it?

Let '[*HIV*]' stand for the fact that Jean has HIV and let '$\Delta_1$' stand for a full explanation of why Jean developed AIDS of which [*HIV*] is a part. Now let '[*Transfusion*]' stand for the fact that he had an infected blood transfusion and let 'Γ' stand for a full explanation of [*HIV*] that includes [*Transfusion*]. Finally, let '$\Delta_2$' stand for the set obtained by substituting Γ for [*HIV*] in $\Delta_1$.[12] My suggestion is this: $\Delta_2$ is a full explanation of why Jean had AIDS.

That is, if we substitute the full explanation of why Jean had HIV, Γ, for the fact that he had HIV, [*HIV*], into the full explanation of why he developed AIDS, $\Delta_1$, that produces another full explanation of why he had AIDS, $\Delta_2$. Now, $\Delta_2$ is clearly not identical to $\Delta_1$, but it is also not genuinely different from it. Moreover, since [*Transfusion*] is included in $\Delta_2$, and $\Delta_2$ is a full explanation of why he developed AIDS, [*Transfusion*] is a partial explanation of why Jean developed AIDS.

To answer the question then, in 'plain' English: the fact that Jean had an infected blood transfusion is a part of the full explanation of why he developed AIDS that is got by substituting the full explanation of why Jean had HIV for the fact that he had HIV into the more proximal full explanation of why he developed AIDS.

# 6 Conclusion

I have argued that the exclusion principle is false and that seemingly excluded facts can explain an *explanandum* by explaining something that, in turn, explains that *explanandum*. In the next chapter I will set out the principles of my indirect theory of normative reason explanation.

[12] That is: $\Delta_2 = \Gamma \cup \Delta_1 \setminus \{[HIV]\}$.

# (XII)

# Explaining why we act

> *In which I say how normative reasons (and the appearance of them) explain why we act. I suggest that normative reasons explain an agent's action by explaining their belief that, in turn, explains the agent's action. I suggest that they explain an agent's belief by explaining the appearance of them that, in turn, explains the agent's belief. I set out the implications of this view for explanatory rationalism and for anti-psychological theories of reasons more generally.*

In light of the remarks of the previous chapter, my proposed answer to the question of how normative reasons explain our actions is perhaps clear:

> Facts about an agent's [normative][1] reasons explain an agent's actions whenever they explain why she has the (true) beliefs she has about her [normative] reasons, beliefs that in turn explain her actions. (Smith 1998, 38)

This is the indirect theory of normative reason explanation. According to this theory the fact that my friend won an award explains why I congratulated her because it explains why I believed that she had won an award, which, in turn, explains why I congratulated her.

The indirect theory of normative reason explanation has been variously considered, endorsed or rejected, by others, though mainly in passing.[2] However, I do not think it has been considered as thoroughly as it ought to have been, because, I suggest, its implications for what an agent's reason for acting could be are profound (regardless of whether or not one accepts explanatory rationalism).

Moreover, as I will show, the indirect theory applies equally well as an account of how experiences explain actions – that is, an experience explains an agent's action if it explains the belief that explains the agent's action.

In the next chapter I will consider what the indirect theory should say about how normative reasons and the appearance of them explain why it is rational to act.

---

[1] These are 'justifying' reasons in Smith's original – the substitution I make is purely terminological.

[2] See: Collins (1997, 111), Smith (1998, 38), Dancy (2000, 109–101), Davis (2005, 74–75), Saporiti (2007, 306), Raz (2009, 197) and Gibbons (2010, 359). Dustin Locke (2015) gives a more thorough treatment of a theory of this kind, however, Locke's treatment differs significantly from mine, particularly in so far as it is focused on what it is for something to be an agent's reason for acting.

# 1 When normative reasons explain

The principle behind the indirect theory is that when an agent acts for a reason 'there is the "proximal" explanation of the action, given by specifying the psychological state of the agent. Then there is the "distal" explanation of the action, given by specifying what is responsible for the agent getting into that state.' (Dancy 2000, 109) For this to be a fruitful account of how normative reasons explain actions, it must be the case that a normative reason can be 'what gets' an agent into the state of believing that normative reason.

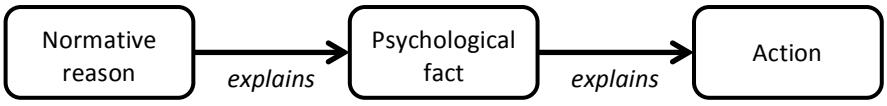


**Figure XII-1:** The indirect theory of normative reason explanation

But how does a fact about the world explain one's belief in it? That explanatory relation is, I suggest, also indirect.

## 1.1 How normative reasons explain beliefs in them

Why do I believe that the earth is spherical? Because it is spherical! That which I believe explains why I believe it. Likewise: if I'm being rained on, and I'm jumping to avoid the puddles and I see others rushing for shelter, if you were to then ask me why I believe it's raining I will reply, incredulously, 'Because it is raining!'

How could the fact that it is raining explain why I believe that it is raining? Perhaps you think along these lines: it's not the fact that it was raining that explains why I believed that it was raining, it is only the fact that it appeared to me as though it was raining. If it weren't raining but it still appeared to me as though it was, I would still have believed that it was. And so on. This view, as I argued in § (IX)5, is just another instance of The Explanatory Exclusion Problem, which the previous chapter discredited. In short: even though how things appear to be intermediate between the world and our beliefs about it, that does not prevent the world from being able to explain those beliefs.

To wit: the fact that things appear to be a certain way may be the more proximal explanation of an agent's belief that they are that way. However, if things actually are that way, and they appear to be that way *because* they are that way, then, given the transitivity of the explanatory relations involved, the fact that they are that way can explain why an agent believes them to be that way. This is, I suggest, an *indirect* theory of *belief* explanation.

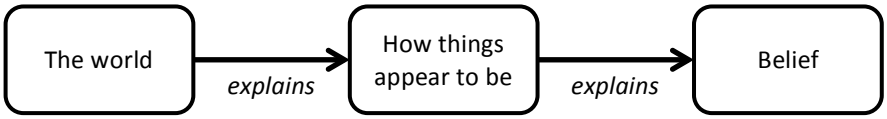


**Figure XII-2:** The indirect theory of belief explanation

Suppose that I believed that it was raining because I looked out of the window and saw rain. Why did it appear to me as though it was raining? Well, in part, *because it was raining*. The fact that it was raining explains why I saw rain and thus why it appeared as though it was raining, which explains why I believed that it was raining. So, because the explanatory relations involved are transitive, the fact that it was raining explains why I believed that it was raining.

If you doubt that the explanatory relations involved are transitive, please suspend your doubts for the moment: this is a subject I will discuss at length in subsequent chapters.

Likewise, I believed that my friend had won an award, in part, because I read that she had won an award. And I read that she had won an award, in part, because she had won an award.[3] The fact that she had won an award explains why I read that she had won an award, which explains why I believed that she had won an award. Again, because the explanatory relations involved are transitive, we can say that the fact that my friend won an award explains why I believed that she had won an award.

So, the way that normative reasons explain our beliefs in them is by explaining why it seemed to us as though those normative reasons were the case.

## 1.2 How normative reasons explain actions

What the indirect theory of normative reason explanation suggests is this: I congratulated my friend, in part, because I believed that she had won an award, and I believed that she had won an award, in part, because I read that she had won an award, and I read that she had won an award, in part, because she won an award. So, because the explanatory relations involved are transitive, we can say that I congratulated my friend, in part, because she won an award.

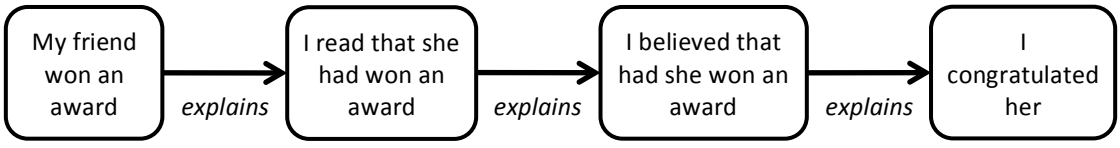


**Figure XII-3:** The explanation of why I congratulated my friend

[3] We could spell out the chain further: I read that she had won an award, in part, because the newspaper printed an article about it, and they printed an article about it, in part, because a reporter wrote an article about it, and they wrote about it, in part, because they witnessed her win the award and they witnessed her win it, in part, because she won it. The point is that given the transitivity of the explanatory relations involved there is no requirement to spell out these intermediate steps.

## 1.3 When normative reasons don't explain

Recall that we don't tend to give normative reason explanations in instances of false beliefs or in Gettier cases – we say that, in those cases, the normative reason does not explain the agent's action. The indirect theory of normative reason explanation tells us why: because in both false belief cases and Gettier cases, the normative reason does not explain the agent's belief in it.

First, a false belief case: when Sally ran because she (mistakenly) believed that a bear was chasing her, we don't say that she ran *because* a bear was chasing her. Why not? Because the proposition *that a bear was chasing her* does not explain why she believed that a bear was chasing her. Clear enough.

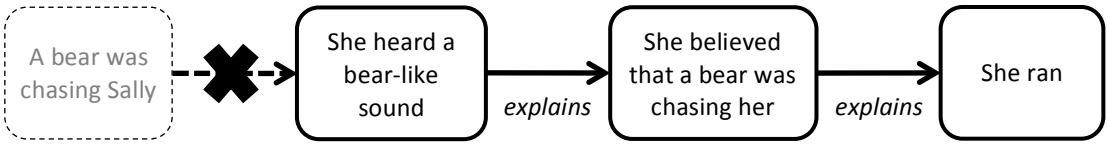


**Figure XII-4:** The explanation of why Sally ran

Second, a Gettier case: recall that Edmund's normally reliable friend told him, on a whim, that the ice in the middle of the lake was thin, although she had no idea about the actual status of the ice. As a result Edmund skated by the edge of the lake; that is, he skated by the edge of the lake, in part, because he believed that the ice in the middle was thin.

Now, as it turns out, it actually was thin. However, as we have already acknowledged (see § (IV)1.4), the fact that the ice in the middle of the lake was thin does not explain why Edmund skated at the edge. Why not? Well according to the indirect theory tit is because it does not explain why he *believed* that it was thin; and it does not explain why he believed that it was thin because it does not explain why his friend told him that the ice was thin (which is what explains why he believed that it was thin).

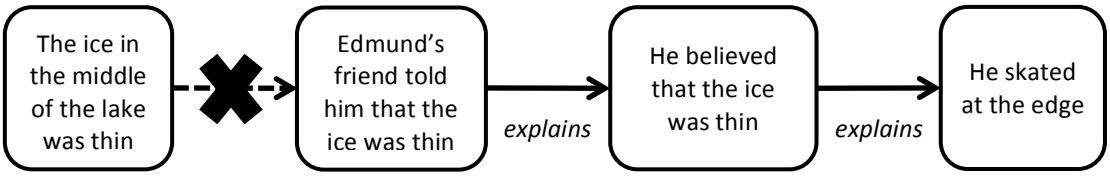


**Figure XII-5:** The explanation of why Edmund skated at the edge of the lake

# 2 Implications for explanatory rationalism

Recall that explanatory rationalism requires that an agent's reason for acting must both explain their action and explain why it was *pro tanto* rational for them to do it. As I set out in

§ (VI), if explanatory rationalism is to be consistent with the *prima facie* reasonable claims set out in §§ (II)-(IV), the following must be true:

(R1) I congratulated my friend because she had won an award.

(R2) I congratulated my friend because I read that she had won an award

(R3) It was *pro tanto* rational for me to congratulate my friend because she had won an award.

(R4) It was *pro tanto* rational for me to congratulate my friend because I read that she had won an award.

The discussion of the previous section has demonstrated how it is that (R1) and (R2) are true: the *explanans* in each statement explains why I congratulated my friend by explaining why I believed that she had won an award.

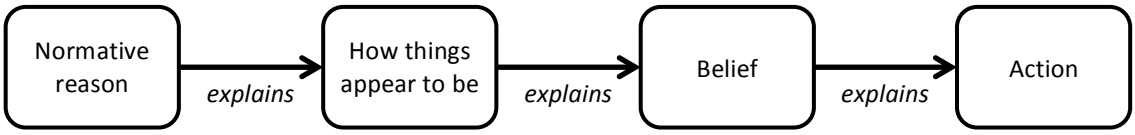


**Figure XII-6:** Explaining why we act

In the next chapter I will discuss (R3) and (R4), but before then I want to discuss the implications of the indirect theory of normative reason explanation for anti-psychological theories of reasons more generally.

# 3 Implications for anti-psychological theories of reasons

Recall that The Explanatory Exclusion Problem was the motivating argument for psychologism about the reasons for which we act. The reasoning went like this: an agent's reason for acting must always explain their action (recall § (IV)1.2), however, only features of an agent's psychology can explain their actions (as per The Explanatory Exclusion Problem), therefore, an agent's reason for acting must always be a feature of their psychology.

This argument is, as we have noted, the major argument against any theory that suggests that reasons are sometimes not features of our psychology.[4] The rejection of the exclusion principle undermined that argument, strengthening the case for anti-psychological theories.

The indirect theory of normative reason explanation now provides us with an account of why it is wrong: normative reasons can explain actions, and they do so by explaining the beliefs that

[4] Note: anti-psychological theories don't insist that reasons are never features of our psychology, just that they *at least sometimes* aren't.

explain our actions. This theory does not deny the primacy of the psychological, on which psychologism insists, but it nonetheless provides a role for the world in explaining what agents do. The indirect theory of normative reason explanation thus provides a response to psychologism for all anti-psychological theories, and not merely explanatory rationalism.

# 4 Conclusion

One of the challenges I set myself was showing how it was that normative reasons could explain an agent's action. I take that challenge to have now been met. In particular, I have argued that, regardless of whether or not you accept explanatory rationalism (and the other arguments for it that follow), the indirect theory of normative reason explanation can still provide you with an account of how it is that normative reasons can explain our actions.

The next chapter discusses how it is that normative reasons (and perceptual experiences and the like) can explain why it is rational for an agent to do something. Before that discussion, in the Appendix to this chapter, I consider two potential objections to the indirect theory of normative reason explanation.

# Appendix

## A.1 Objections

**Objection 1** You have suggested that a normative reason explains an agent's action only if it explains a feature of their psychology that explains their action. However, here is a counterexample to that claim:

> That it is about to rain may explain why everyone is coming in. Is their belief that it is about to rain to be explained by its being about to rain? Or is it rather the blackness of the clouds and the sudden drop in temperature? These are not themselves to be explained by its being about to rain. The clouds are not black *because* it will shortly rain. (Dancy 2000, 112)

For expositional simplicity, let's restrict Dancy's example to claims about someone in particular; call him 'Jim'. For Jim, the argument goes like this:

(a) The fact that it is about to rain is a normative reason for Jim to come in;

(b) The fact that it is about to rain explains why Jim is coming in; and

(c) The fact that it is about to rain does not explain why Jim believes that it is about to rain.

If (a), (b) and (c) are all true then it is not true that, as I have argued, a normative reason explains an agent's action only if[5] it explains a feature of their psychology that in turn explains their action.[6]

**Response** I am happy to agree that (a) is true, so I must reject either (b) or (c). There are some[7] who reject (b), on the ground that facts about the future can't explain present actions. I have some sympathy for that response; however, it may not fare so well with facts known from inference, and may rely on a causal analysis of the explanatory relations involved. Regardless, to my mind there is a more compelling response: I do not know how one can consistently maintain both (b) and (c).

The only reason that Dancy gives in defence of (c) is that the blackness of the clouds etc. is not explained by the fact that it is raining. That is true enough. But to make that argument is to embrace an indirect theory of belief explanation – namely that some fact *p* explains the fact that an agent believes that *p* only if it explains some fact that, in turn, explains the fact that *p*. That is, Dancy presupposes the indirect theory of belief explanation in suggesting that the fact that it is about to rain can't explain why Jim believes that it is about to rain because it does not explain the blackness of the clouds etc. My concern is that I don't know why we should be indirect theorists about belief explanation but not about action explanation.

Of course, if one were to deny (c), and insist that the fact that it is about to rain does explain why Jim believes that it is about to rain, then one might wonder what the explanatory connection between those two facts is (given that it is unmediated by any perceptual experience). However, I don't see how one can wonder this without likewise wondering what the explanatory connection between the fact that it is about to rain and the fact that Jim came in is. My point is this: I don't see what basis one could have for thinking that both (b) and (c) are true – either one accepts (b) and, *for the same reason*, rejects (c); or one accepts (c) and, *for the same reason*, rejects (b).

So, if there is no consistent basis for the truth of both (b) and (c), then this is no objection to the indirect theory of normative reason explanation.

---

[5] At least in non-weird cases.

[6] An argument of this kind can likewise be made for facts that are believed on the basis of inference (i.e. where a perceptual experience that explains one's belief that *p* is also seemingly is not explained by the fact that *p*). Alvin Goldman (1967) considered these sorts of examples in his causal theory of knowledge, which is similar to the indirect theory in several respects. However, his response to them differs from mine.

[7] For instance: Davis (2003, 456), Gibbons (2010, 359) and Locke (2015, 194).

**Objection 2** Even if we accept this indirect theory of normative reason explanation, it does not guarantee that you congratulated your friend because she won an award (i.e. it does not guarantee the truth of (R1)). Why not? Because one may be sceptical as to whether or not the fact that your friend won an award really explains the fact that you believed that she had.

For instance, suppose that your friend's award is particularly obscure and you just happen to stumble upon a small article about it in a newspaper that you wouldn't normally read. Under these circumstances then (at least on, for instance, a difference-making account of explanation) one might say that the fact that your friend won an award does not really explain why you believed that she had,[8] and it consequently does not explain why you congratulated her. Thus, in spite of your indirect theory of normative reason explanation, (R1) may still be false.

**Response** I have three responses to this objection. First, I could just accept that (R1) is not true when my discovering about my friend's award is so chancy. There might be other examples (e.g. my friend wins a Nobel Prize), for which the explanatory connection is sufficiently robust to counterfactuals for the fact that my friend won an award to count as an explanation of my action. One strategy could thus be to just restrict (R1) to a claim about such cases. There would be little lost for my theory in making such a restriction.

My second response, however, is to note that even if my friend's award had been obscure, and I had only found out about it because I stumbled upon an article about it in a newspaper I wouldn't normally read, I still think that the fact that my friend won the award (partly) explains why I believed that she did and, therefore, why I congratulated her. So, to the extent that your account of explanation implies that it doesn't, it is not really the sort of account that I am anyway inclined to accept.[9]

[8] Since there are very nearby possible worlds in which you don't see the article, but in which she nonetheless wins the award.

[9] What may be at work here is an explanatory analogue of Ned Hall's (2004) claim that there are two concepts of causation: *productive* and *counterfactual dependence*. I suggest that difference-making accounts of explanation can be understood as particular kind of explanatory analogue of counterfactual dependence concepts of causation (even without restricting the discussion only to causal explanation). Now, whilst I don't want to reject difference-making accounts of explanation, I would suggest that there is a *bona fide* manner of explanation that is the explanatory analogue of Hall's productive causation. The relevance of this observation to this discussion is that Hall demonstrated that one can have productive causation without counterfactual dependence, and vice versa. Thus, my suggestion (modestly made) is that there may be two different concepts of explanation that can likewise come apart. Thus, it may be that even though the fact that my friend won an award does not explain why I believed that she had if we have in mind a particular sort of *counterfactual dependence* concept of explanation (i.e. difference-making), it may nonetheless still explain it if we have in mind the *productive* concept.

My third response: in the chapters that follow I argue that if an agent knows that *p* then that entails that there is a particular explanatory relation between the fact that *p* and the fact that they believe that *p*. If you accept this, then you must either reject the claim that I knew that my friend had won an award or accept that the fact that my friend had won an award explains why I believed that she had.[10] Since I *do* think that (even in the chancy case) I knew that my friend had won an award, I am also inclined to think that the fact that she had won an award explains why I believed that she had, and, thereby, explains why I congratulated her. However, if, in the chancy case, you don't think that I knew that my friend had won an award, then you could (as we did for the first response) just restrict (R1) to a claim about cases in which I knew that my friend won an award, without much loss for my theory.

[10] If you reject the claim about knowledge that I go on to make you are, of course, under no such compulsion. However, if you reject this claim then you will perhaps have more significant objections to my theory than just this.

# (XIII)

# Explaining why it is rational to act

*In which I say when something explains why it's rational to act, and when it doesn't. I suggest that normative reasons or appearances explain why it is rational to act only if they explain those beliefs that in turn explain why it is rational to act. I note that it is tempting to infer that if an agent's belief explains why it is rational for them to do some action then whatever explains that belief also explains why it is rational for them to do that action. I show how that inference leads to an apparent dilemma for explanatory rationalism. I counsel against that inference, by noting that different kinds of explanatory relations may not be transitive with each other. I then set out the task ahead: showing that the explanatory relations concerned are transitive when, and only when, explanatory rationalism needs them to be.*

There is a seemingly clear way in which the account of the previous chapter could also be used as an account of how it is that normative reasons (or the appearance of them) explain why it is rational for an agent to act: we can say that a normative reason (or the appearance of one) explains why it is rational for an agent to act if it explains the belief that explains why it is rational for the agent to act.

So, for instance, when I look out of the window and see rain, we can say that the fact that it is raining explains why it is rational for me to take my umbrella *because* it explains why I believed that it was raining. This proposal vindicates explanatory rationalism's claim that things other than features of an agent's psychology could explain why it is rational for them to do something.

However, as I will argue, we should not assume that just because some fact explains an agent's belief and their belief explains why it is rational for them to do some action, that that fact also explains why it is rational for them to do that action. That is, I suggest, there are occasions on which the explanatory relations involved aren't transitive.

In the next chapter I will provide an account of what separates those cases in which the explanatory relations involved are transitive from those in which they aren't.

## 1 Another indirect theory

Recall that I need to demonstrate the following:

(R3) It was *pro tanto* rational for me to congratulate my friend because she had won an award.

(R4) It was *pro tanto* rational for me to congratulate my friend because I read that she had won an award.

The question of how these claims could be true is an instance of a more general question, namely: how is it that either normative reasons (*qua* facts about the world) or the appearance of them could explain why it is (*pro tanto*) rational for an agent to act? My answer is that they explain indirectly.

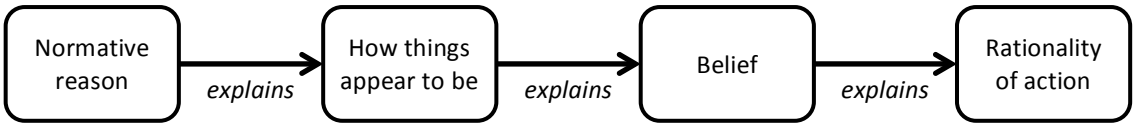


**Figure XIII-1:** Explaining why it is rational to act

The fact that it is raining explains why it appeared to me as though it was raining, which explains why I believed that it was raining, which explains why it was rational for me to take my umbrella. So, because the explanatory relations involved are transitive, the fact that it was raining explains why it was rational for me to take my umbrella.

Likewise, as we established in the previous chapter, the fact that my friend won an award explains why I read that she had won an award, which explains why I believed that she had won an award, which explains why it was *pro tanto* rational for me to congratulate her. And this is why (R3) and (R4) are true: we can say that it was *pro tanto* rational for me to congratulate my friend both because she won an award and because I read that she had won an award, and we can say that because the explanatory relations involved are transitive.

## 2 Is explanation transitive? An apparent dilemma

I have said that normative reasons and how things appear to be can explain why it is *pro tanto* rational for an agent to do some action only if, and *because*, they explain one of the agent's beliefs, which, in turn, explains why it is *pro tanto* rational for them to do that action.

As I have noted, this indirect manner of explanation relies on the transitivity of the explanatory relations involved. With that in mind, we might then reason from this observation to the following theory about what explains why it is rational to act:

- *The Naïve Theory:* if the fact that *A* believes that *p* explains why it is *pro tanto* rational for *A* to *φ* then *whatever* explains why *A* believes that *p* also explains why it was *pro tanto* rational for *A* to *φ*.

In what follows I will argue that the naïve theory is false. Before I do so, however, I want to show how the naïve theory might seem to create a dilemma for explanatory rationalism, irrespective of whether it is true or false.

Recall what explanatory rationalism has to say about the reasons for which we act:

- *Explanatory rationalism about the reasons for which we act*: For any *p*, *p* is *A*'s reason for *φ*ing if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and explains (in the right way) why *A φ*'d.

This, combined with the naïve theory, means that if *A φs* because *A* believed that *p* and it was *pro tanto* rational for *A* to *φ* because *A* believed that *p*, then *whatever* explains why *A* believed that *p* is also *A*'s reason for *φ*ing.

But notice that sometimes we can explain why someone believed something by citing things that are plainly not the reasons for which they act, such as brain aneurisms or psychosis. Were the naïve theory true, then, the explanatory rationalist would be forced to conclude that, if an aneurism explains why an agent has some belief, and they act rationally on that belief, then their reason for acting is that they had a brain aneurism. This is an undesirable conclusion.[1]

And yet, if the naïve theory isn't true, then explanation is not a generally transitive relation. But the indirect theory of normative reason explanation, and, more generally, my account of why (R1) to (R4) are true, relies on the transitivity of explanation.

Here, then, is the apparent dilemma: either (i) we accept the naïve theory, which means rejecting explanatory rationalism about reasons for which we act;[2] or (ii) we reject the naïve theory by rejecting the transitivity of explanation, which means giving up on the truth of (R1) to (R4).[3] The next section is devoted to arguing that this dilemma is only apparent and that the appearance of a dilemma is the result of a mistake – the same mistake that led to the naïve theory.

[1] Of course, the fact that one has a brain aneurism or psychosis could be a reason for which one does something (e.g. goes to the doctor). However, if that brain aneurism just causes one to have an irrelevant belief through some arational process, we would typically not want to conclude that when an agent acts because of that belief, their reason for acting is that they had a brain aneurism.

[2] Or saying that when a brain aneurism causes me to believe that it is raining that my reason for taking my umbrella was that I had a brain aneurism. I take this to be worse than rejecting explanatory rationalism.

[3] Which means, in turn, that explanatory rationalism reduces to psychologism. It reduces to psychologism because, if explanation can't be transitive, then only features of an agent's psychology can explain why they act or why it is rational for them to act.

# 3 The apparent dilemma is not a dilemma

One simple, but mistaken, way to arrive at the dilemma is to think that there is only one kind of explanatory relation. Were that true then any instance of a failure of transitivity (i.e. if the naïve theory were false) would mean that explanatory relations are not transitive.[4] However, as the discussions of previous chapters have suggested, there is not just one kind of explanatory relation. We can at least distinguish causal from non-causal explanatory relations, and I suggest that there are probably still more fine-grained distinctions between explanatory relations that that distinction ignores.

A more nuanced but, as I will argue, equally mistaken approach is to maintain that although there are different kinds of explanatory relation, if any explanatory relations are transitive then they are all transitive with each other. I suggest that it is this fallacious assumption that led to the naïve theory.

My argument against this assumption proceeds in two stages: firstly I will demonstrate that some explanatory relations are transitive; secondly I will give an example in which explanatory relations are not transitive. Together these amount to a counterexample to the claim that if any explanatory relations are transitive then they all are.

## 3.1 Some explanatory relations are transitive

Firstly, as § (XI) demonstrated, we have good reasons to think that some explanatory relations are transitive: if they weren't then *distal* explanations of some *explanandum* would never really be explanations of that *explanandum*.[5] Since, in both ordinary and scientific life, we are mostly interested in somewhat distal explanations, and since the explanations we normally give are *somewhat* distal, our ordinary and scientific explanatory practice assumes that at least some explanatory relations are transitive. So, I will take it as a given that some explanatory relations are transitive, because that is the best account of why distal explanations that involve such relations explain.

[4] In so far as a chain of explanatory relations does not guarantee an explanatory relation linking the two explanations of the chain.

[5] Of course it's possible that *some* distal explanations explain for reasons other than the transitivity of the explanatory relations involved. I'm not sure what those reasons could be, but one might be able to construct examples. My point is rather that the transitivity of the explanatory relations involved is really the best and simplest account of how distal explanations explain, and it really does explain how they explain in at least some cases.

## 3.2 Not all explanatory relations are transitive with each other

In this section I will provide an example in which the transitivity of explanation fails, and I will suggest that it is because the explanatory chain involves different kinds of explanatory relations that aren't transitive with each other.

### 3.2.1 When explanation isn't transitive

Recall that Sally believed that a bear was chasing her because she heard a bear-like noise. Why did she hear a bear-like noise? Because the wind rustled the trees in such and such a way. That is, the noise that Sally heard had nothing to do with any bear, but it nonetheless sounded very bear-like.

We have already noted that it was *pro tanto* rational for Sally to run because she believed that a bear was chasing her. And what explains her belief that a bear was chasing her was that she heard a bear-like noise (it appeared to her as though a bear was chasing her), and what explains that was the fact that the trees rustled (in the way that they did). Thus, there is a chain of explanatory relations from the fact that it was *pro tanto* rational for Sally to run to the fact that the trees rustled. But should we say that it was *pro tanto* rational for Sally to run (even partly) because the trees rustled?

I don't think we should. It seems as though even if one were to give this sort of explanation one would then be forced to add that Sally believed that a bear was chasing her. And, I suggest, that is because it is only really the belief that is doing the explaining.[6]

However, if the fact that the trees rustled does not (even partially) explain why it was *pro tanto* rational for Sally to run,[7] then the explanatory relations involved in this case must not be transitive.[8]

Perhaps you object: this does not violate transitivity because the odd rustling of the trees does not explain why she heard a bear-like sound. Of course, you say, it explains why she heard *something*, but it doesn't explain why that sound sounded like a bear. While this would be all the better for me if I agreed with it, unfortunately I don't: of course other factors may do more to explain the particular character of the noise made, but the fact that the trees rustled in that way is, I suggest, *a part* of the explanation of why she heard a bear-like noise.

---

[6] If you worry that this is too close to the elliptical view that I rejected, note how we *aren't* forced to add anything about the belief when we say that I took my umbrella because it is raining.

[7] Of course, if your view of explanation is that it is *purely* a relation of counterfactual dependence, then it does explain it. However, I think that is a wrong account of explanation (see my response to Objection 2 in the Appendix to the previous chapter).

[8] This follows necessarily.

### 3.2.2 Why explanation wasn't transitive

The rustling in the trees explained the noise that Sally heard, the noise explained her belief, and her belief explained why it was rational for her to run. However, I've suggested, the explanatory relations here[9] aren't all transitive with each other. Why not? Because, I suggest, they are different sorts of explanatory relation.

What sort of explanation is the explanation of why it was *pro tanto* rational for Sally to run? Well, it isn't causal. Even if we allow that our beliefs cause our *actions*, it's still not the case that they *cause* it to be rational for us to act – any more than the fact that a person is in trouble *causes* it to be right to save them. Causation just seems to be the wrong concept to invoke when describing this sort of explanatory relation. That is, the explanatory relation between my belief and the rationality of my action is a *non-causal* explanatory relation.

In contrast, the only sense in which the rustling of the trees explains why Sally heard a bear-like sound is a causal one; that is the explanatory relation here is a causal explanatory relation.

So, the explanatory relations between (i) the rustling trees and the noise Sally heard and (ii) her belief and the rationality of running are clearly of two different kinds. And I suggest that the different character of the explanatory relations involved is why they aren't transitive with each other. Note that I'm not saying that causal explanatory relations are never transitive with non-causal explanatory relations[10] – all I am saying is that *these* particular sorts of causal and non-causal explanatory relations aren't transitive.

## 3.3 The apparent dilemma isn't a dilemma

I said that the naïve theory and the apparent dilemma it produces were (at best) the result of the assumption that if any explanatory relations are transitive then all explanatory relations are transitive with each other. The argument of the previous two sections disproves that assumption – while some kinds of explanatory relations are transitive, it does not follow that different kinds of explanatory relations are transitive with each other.

Thus, there is no dilemma for explanatory rationalism because the naïve theory and the consequent apparent dilemma were based on the mistaken assumption that all explanatory relations must be transitive with each other.

[9] That is: between (i) the trees and the noise; (ii) the noise and the belief; and (iii) the belief and the rationality of running.

[10] That is, there may be other kinds of causal and non-causal explanatory relations that are transitive with each other.

However, explanatory rationalism is not out of difficulty yet, in the next section I will discussion a new challenge posed by this response. Namely this: why should it be that the explanatory relations involved are transitive whenever explanatory rationalism needs them to be, and not when it needs them not to be?

# 4 The challenge

We are talking about three sets of explanatory relations: the relation from the world to an appearance, from an appearance to a belief, and then the relation of that belief to the rationality of an agent's action. What I seem to be saying about these explanatory relations is this: sometimes they are transitive with each other, and sometimes they aren't. In particular, I am saying that the explanatory relations in Figure XIII-2 are transitive, while the explanatory relations in Figure XIII-3 aren't.

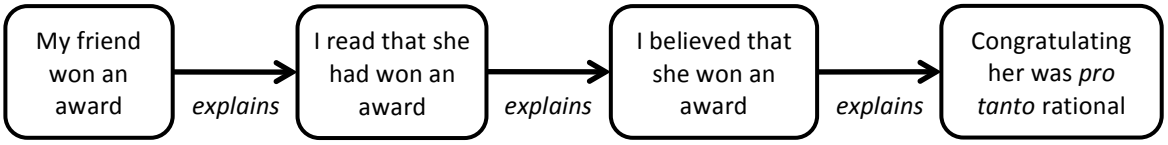


**Figure XIII-2:** A chain of explanatory relations that are all transitive

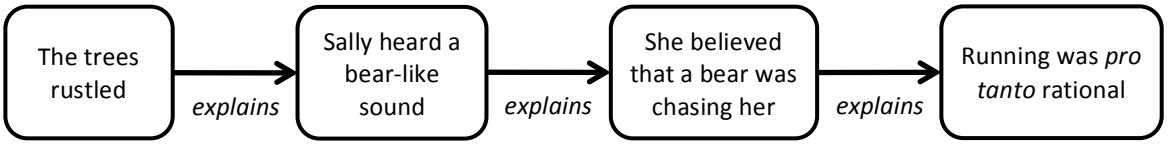


**Figure XIII-3:** A chain of explanatory relations that are not all transitive

Isn't this just ad-hoc? Why should we suppose that the relations are transitive when explanatory rationalism needs them to be, but aren't when it needs them not to be? That is the remaining challenge for explanatory rationalism.

I want to consider one response to this challenge that is natural, but won't work, before setting the stage for my own response.

# 5 The unsuccessful natural strategy

A natural strategy might be to reason as follows: perhaps in the second case the failure of transitivity is not in the explanatory relation *per se*, but in the putative *explanans* (i.e. the rustling of the trees); it's neither the content of Sally's belief nor is it something that her belief was based on – it's of no epistemic importance. And perhaps because it's of no epistemic importance, it's just the wrong sort of thing to do the right sort of explaining. I could then refine my theory in a way that excluded these sorts of (epistemically unimportant) putative

*explanans* for some principled reason, and I could then insist that whatever met those conditions *and* explained my belief thereby also explained what the belief explained.

There are two problems with this strategy: firstly, it is not at all clear to me that one could make such a restriction without it still being, in some respect, ad-hoc. Secondly, and more seriously, explanatory relations may fail to be transitive even when the putative *explanans* is a normative reason (i.e. the right sort of thing), as the next example demonstrates.

### 5.1.1 Another example of failure of explanatory transitivity

Suppose that I am in a sealed room that, unbeknownst to me, is slowly being filled with carbon monoxide. Suppose further that, after a short while, the carbon monoxide causes me to hallucinate, and, by sheer chance, I have a hallucination of a reliable friend bursting into the room and warning me that it is filling up with carbon monoxide. It is rational for me to leave,[11] and I duly do so.[12]

The fact that the room was filled with carbon monoxide (partly) explains why I had the hallucination. Of course, it is only a part of that explanation – other factors of my psyche and the like will do more to explain *the content* of my hallucination, but the fact that there was carbon monoxide in the room is, I suggest, a part of the full explanation of why I had that hallucination.

Moreover, the fact that I had that hallucination explains why I believed that there was carbon monoxide in the room. And the fact that I believed that there was carbon monoxide in the room explains why it was *pro tanto* rational for me to leave.

Again, we have an explanatory chain connecting the fact that there was carbon monoxide in the room with the fact that it was *pro tanto* rational for me to leave. But does the fact that there was carbon monoxide in the room explain why it was *pro tanto* rational for me to leave? I don't think so.

It's tempting to put it like this: I wasn't really aware of that fact. I am not really responding to the fact that there is carbon monoxide in the room because my belief is just true by happy accident. That is, that fact just seems to stand in the wrong sort of relation to the fact that I believed that it was raining, i.e. it stands in a *purely* causal explanatory relation to that fact.

[11] Assume, if you like, that I know that I am not prone to hallucinations, and also don't believe that there is anything that might make be hallucinate now.

[12] Thanks for this example are owed to John Roberts, who put it to me in a seminar. This is what one might call a 'deviant' Gettier case, involving both a justified, true belief that falls short of knowledge and a deviant causal chain linking the fact itself with the belief. This sort of case is a well-established problem for Goldman's (1967) causal theory of knowledge (see e.g. McDonnell 2015). I will discuss deviant causal chains in greater detail in the next chapter.

### 5.1.2 The natural strategy is unsuccessful

Now we can see why the natural strategy is unsuccessful: in the example just considered, the would-be *explanans* is the right sort of thing – it is a normative reason, indeed, it is actually what I believe – and yet the explanatory relations involved still aren't transitive. So, as a way of distinguishing between the different cases, the natural strategy fails. That is, it's not because of the nature of the would-be *explanans* that transitivity of explanation fails.

The next section sketches out my strategy, which will be developed in detail in the next chapter.

# 6 The mysterious strategy

Here is the challenge: I need a principled account of why it is that the explanatory relations from the believed fact to the rationality of the action are transitive in the 'award' case but not in either the 'Sally' case or the 'carbon monoxide' case. What is that account?

I've hinted at my answer already, and McDowell expresses the intuition about what marks out the 'award' case from the others nicely thus: 'we can say that the fact itself is exerting a rational influence on the agent's will; we can say that… the agent is responding rationally to the fact itself.' (2013, 17) Now, in § (X)4.2.1, I lamented the inscrutably metaphorical character of these sorts of remarks, so, whilst they will do as an expression of the intuition, I would be falling well short of my own standard if I left it there.

In the chapters that follow I want to give a non-metaphorical characterisation of what it is that distinguishes the 'award' case from the others. In particular, I will argue, there is a *mysterious*, transitive, non-causal explanatory relation between the fact that my friend won an award and the fact that I believed that she did which is lacking in the other cases.

In the subsequent chapter, I will argue, this *mystery* relation is transitive with the non-causal explanatory relation that obtains between the fact that I believed that my friend had won an award and the fact that it was *pro tanto* rational to congratulate her.

This argument provides a principled reason for saying why the explanatory relations are transitive in the 'award' case but not in the others: because, in the 'award' case there is an explanatory relation between the world and my belief that is lacking in the other cases.

# (XIV)

# The Mystery Relation

*In which I introduce the mystery relation. I suggest that a mysterious, non-causal relation obtains between a belief and the justification that it is based on when that belief is justified. I argue that the mystery relation must be non-causal, because, as deviant causal chains demonstrate, a merely causal relation between a belief and some justification for it is not sufficient for that belief to be justified. I suggest that this exact same mysterious relation relates: the belief that p to the fact that p when the belief that p is knowledgeable; a justification for the belief that p to the fact that p when that justification affords the opportunity for knowledge; and an action to some belief that explains why it is rational when that action is done intentionally. I argue, furthermore, that this mystery relation is a transitive, explanatory relation.*

Here are some questions: what distinguishes a justified belief from a merely justifiable one? What separates a knowledgeable belief from a belief that one holds, when one is in a position to know it, without knowing it? What is the difference between a justification that affords the opportunity for knowledge from one that doesn't? And, lastly, what distinguishes an action done intentionally from a mere bodily movement?

I do not claim to know the answer to these questions. However, in what follows I want to see what can be said *without* offering a theory of the difference between these cases. For instance, the difference between a justified belief and a merely justifiable one is already characterised in terms of the epistemic basing relation – the question is how that relation should be understood. But even without answering that question, we can still say some things about the basing relation. In particular, it is widely agreed that the problem of deviant causal chains frustrates a purely causal analysis of the basing relation, so, I suggest, we can suppose that the basing relation is *not merely causal.* Which is to say, I suggest, that there is a *mysterious* non-causal relation between a belief and that which it is based on, when that belief is justified.

I suggest that this exact same mysterious relation relates: the belief that *p* to the fact that *p* when the belief that *p* is knowledgeable; a justification for the belief that *p* to the fact that *p* when that justification affords the opportunity for knowledge; and an action to some belief when that action is done intentionally. I argue, furthermore, that this mystery relation is a transitive, explanatory relation.

This characterisation of the mystery relation provides the basis for the discussion of the next chapter, in which I say that the mystery relation is transitive with the non-causal explanatory relation that obtains between the fact that I believed that my friend had won an award and the fact that it was *pro tanto* rational to congratulate her.

I offer the following analysis modestly. There is no sense in which I take myself to have solved the problem of deviant causal chains for the contexts considered. I also do not think I have offered much of an analysis of what differentiates deviant cases from non-deviant cases other than *something* differentiates them, and that whatever it must be must be in some sense non-causal, explanatory and transitive. I take these observations to be relatively anodyne, although some may disagree with them. My hope is that even these bland observations will suffice for the purposes of the next chapter.

# 1 The mystery relation and justified belief

There is a commonly recognised distinction between a justifi*ed* belief and a merely justifi*able* belief. Here is a typical example:

> A justifiable belief is one the believer could become justified in believing if he just put together in the right way what he already believes. To illustrate, a woman might have adequate evidence for believing that her husband is unfaithful to her, but systematically ignore that evidence. However, when her father, whom she knows to be totally unreliable in such matters and biased against her husband, tells her that her husband is unfaithful to her, she believes it on that basis. Then her belief that her husband is unfaithful is unjustified but justifiable. (Pollock and Cruz 1999, 79)

The woman in this example has some justification for believing that her spouse has been unfaithful (we don't know what it is), which she ignores. In spite of ignoring the justifications she has for believing it, she still ends up forming the belief that her husband has been unfaithful, but bases it on the fact that her father told her that her husband was unfaithful, which is not a justification for believing it (because her father is known to be biased and unreliable).

Her belief is merely *justifiable*, and not *justified*, because a justifiable belief is a belief for which one has some justification (which she does), but a justified belief is one that is *based* on a justification one has for it (which hers is not).

As an aside: I am straying into epistemology here. To restrain the bounds of my assertions, let me state my assumptions plainly: all that is meant by a 'justification' here is something that could explain why it is justifiable for one to believe that *p* and upon which one's belief that *p* could be based. What I am calling 'justification' is more typically called a 'reason for belief' in the literature, but I avoid the 'reasons' terminology to avoid conflating that with the present

discussion.[1] Moreover, I will take it as a given that appearances could be justifications for belief.[2]

## 1.1 The epistemic basing relation

What does it mean for a belief to be based on some justification for it? The *de facto* analysis of the 'basing' relation is a causal one. However, it is widely acknowledged (e.g. Korcz 2015) that the possibility of deviant causal chains between a belief and a justification for it frustrates a purely causal analysis of basing by demonstrating that the fact that a justification for a belief stands in a causal relation to it is not enough to ensure that the belief is justified. Typically one has to qualify a causal analysis by saying that the justification must cause the belief 'in the right way' – but a purely causal analysis of what this 'right way' is, is lacking.

Re-purposing the example above: suppose that Eva sees her husband kissing another woman. Suppose that this is adequate evidence of her husband's infidelity (supplant more compelling evidence if you aren't convinced). She ignores the evidence and carries on believing that her husband is faithful. Her father is out of the picture this time, but suppose, instead, that the stress of ignoring what she has seen (it is a difficult thing to ignore) causes a brain aneurism (in spite of the fact that she does manage to ignore what she has seen) that, by incredible coincidence, causes her to believe that her husband has been unfaithful to her.[3] Believing that her husband has been unfaithful to her, she sues for divorce.[4]

In this case her belief that her husband has been unfaithful is justifiable, and it is caused (albeit, in a roundabout way) by the justification she has for believing it. But it seems wrong to say either that her belief was *based* on the justification that she had, or, indeed, that it was really a justified belief.[5]

So, even if, as is popularly thought, 'the basing relation is at least partly a causal relation,' (Pollock and Cruz 1999, 79) the need to stipulate non-deviancy of the causal chain provides at

[1] Although I think an analogous treatment of reasons for belief is possible, it is beyond the scope of this discussion to provide one.

[2] This will mean excluding the more extreme forms of internalism about epistemic justification (e.g. Davidson 2001d) but, adherents to that view are in the minority, as Littlejohn notes: 'most statist internalists defend the view that experiences constitute our reasons for belief.' (Forthcomming, 4)

[3] We could stress the independence of these events from Eva's point of view: she has ignored the evidence, so when asked why she believes that her husband was cheating on her she won't even cite that fact that she's seen him kissing another woman – perhaps she actually managed to forget it. Who knows what she would say, the important point is that she wouldn't say that she saw him kissing that woman.

[4] Suppose whatever you need to suppose in order to ensure that it was rational for her to do so (the arationality of the way she acquired her belief notwithstanding).

[5] This example is stronger than, for instance, Plantinga's (1993, 69 fn. 8) classic example because what does the causing is also a justification for the belief, whereas in Plantinga's case it is not.

least *prima facie* evidence that that the relation is not *merely* casual. Moreover, in the absence of a compelling, purely causal, solution to the problem of deviant causal chains,[6] I will assume that there is no such solution, and that the *basing* relation, whatever it may be, is not *merely* causal.[7]

Thus, if a belief is based on some justification for it (i.e. if it is justified) then there is a non-causal relation (as well as, perhaps, a causal relation) between the belief and the justification(s) it is based on – let's call it the *mystery* relation.

At this point we should do some ontological housekeeping. The relata of the basing relation are probably not facts – they are, perhaps, beliefs, experiences or events or what have you. In the same vein I think that the relata of the mystery relation proper are also probably not facts. However, it will greatly simplify my formal exposition, at, I think, no cost to my argument, if we treat them as relations between facts.[8]

(M1) For any proposition, *p*, if *A* has a justified belief that *p* then the fact that *A* believes that *p* is *mysteriously related* to some justification for it.

## 2 The mystery relation and knowledge

Now I want to convince you that the same mysterious relation obtains between the fact that *p*[9] and the fact that *A* believes that *p* when an agent knows that *p*. Why should you believe this?

First consider that, as in the case of the basing relation, the possibility of deviant causal chains frustrates attempts to give a purely causal analysis of the relationship between the fact that *p* and an agent's belief that *p* when they know that *p*. In order to maintain a causal theory of knowledge, one must insist that the fact causes the belief *in the right way*.

---

[6] McCain (2012) offers a purely causal solution by, as Korcz (2015) puts it, 'removing the chain' and denying that the basing relation is transitive. However, I share Korcz's concern that this theory fails to capture pre-theoretical accounts of what our beliefs are based on, so I don't find it very compelling. More generally, to the extent that one takes causal relations to be transitive, a purely causal analysis of deviant causal chains will be impossible, assuming that the same causal relation obtains between each link in the chain.

[7] Recent solutions to the problem of deviant causal chains by Hyman (2015) and Sosa (2015) support my contention that relations affected by them are not *merely* causal. They argue that causal relations are still necessary for such relations, but that non-deviancy is only guaranteed by a further *non-causal* relation: 'the manifestation of a competence'.

[8] There need be nothing significant about this move: if the relata of genuine basing relations or mystery relations are the truth-makers of propositions (e.g. Sally's believing that a bear is chasing her to the proposition that she believes it), then the basing relations and mystery relations I mention are merely the counterparts of the genuine relations in an ontology of facts.

[9] Again, some ontological housekeeping: properly speaking it is probably the truth-maker of the fact that *p*, that is related to the belief, but I am again transposing this talk to an ontology of facts for the sake of simplicity.

Let's revisit Eva's case: Eva is in a position to know that her husband has been unfaithful[10] because she is watching him kiss another woman. And his doing that is what causes her to see him doing it which is what causes her to put the effort into ignoring it which is what causes her to have a brain aneurism which is what causes her to believe that he's been unfaithful to her. In this convoluted way, the fact that her husband has been unfaithful causes her to believe that he has been unfaithful at a time when she is in a position to know that he has: she believes it without knowing it, even though she is in a position to know it and even though the fact that he had been unfaithful is what caused her to believe that he had.

Thus, given that a causal relation between the belief that *p* and the fact that *p* is not sufficient for that belief to be knowledge (indeed, even if the agent is in a position to know that *p*) and given that the 'the right way' of some fact's causing a belief in it cannot be analysed in purely causal terms[11] we can say that if an agent knows that *p* then there is a non-causal relation between their belief that *p* and the fact that *p*.

So, even if a causal relation between the fact that *p* and the belief that *p* is required for an agent to know that *p* (a not inconsiderable 'if'), a non-causal relation between the fact and the belief is still also required for the agent to know that *p*.

But why suppose that this non-causal relation between the fact and the belief is the same as the mystery relation that basing relations entail? Because, I suggest, the cases are directly analogous: the distinction between a justified belief and a *merely* justifiable one is analogous to the distinction between knowing that *p* and *merely* believing that *p*[12] when one is in a position to know that *p*.

Indeed, the analogousness between the cases is already recognised in the diagnosis of both these cases *as* cases of deviant causal chains. The simplest explanation of this analogy is that the non-causal relation that is lacking in the *merely* causal cases in each is of a common kind, which I have called the 'mystery relation'. Thus:

> (M2) For any proposition, *p,* if *A* knows that *p* then the fact that *A* believes that *p* is *mysteriously related* to the fact that *p*.

## 3 The mystery relation and opportunities to know

Now I want to convince you that the same mystery relation distinguishes those justifications that afford the opportunity for knowledge from those that do not.

[10] Surely!
[11] Which we assume given that such an analysis seems is both unavailable and anyway implausible.
[12] I.e. Believing without knowing that *p*

We said that Eva was in a position to know that her husband was cheating on her. Sally, in contrast, is not in a position to know that a bear is chasing her (not least because one isn't). Likewise (recalling the example from the previous chapter), when I am in a room that is slowly filling with carbon monoxide, I am not in a position to know that the room is full of carbon monoxide (there are no alarms, no visible warnings etc.). What differentiates Eva from Sally and I with respect to our epistemic positions?

Here is a way to characterise the difference: while we all have a justification for believing that *p*, only Eva has a justification that affords the opportunity for knowledge. Characterising our question in these terms, and generalising it beyond these cases, we can ask: what differentiates a justification for believing that *p* that affords the opportunity for knowledge from one that doesn't? I think it is the mysterious relation that a knowledge-affording justification stands in to the fact that *p* that separates it from a justification that does not afford the opportunity for knowledge.

To start, consider these remarks by McDowell, in his discussion of how perceptual knowledge is possible (given the possibility of illusion):

> Suppose someone is presented with an appearance that it is raining. It seems unproblematic that if his experience is *in a suitable way* the upshot of the fact that it is raining, then the fact itself can make it the case that he knows that it is raining. (McDowell 1982, 474 emphasis added)

What is this *suitable way*? What is the relation between the 'fact itself' and the man's experience that makes it possible for the fact to 'make it the case that the he knows that it is raining'? Well, it isn't *merely* causal.

Suppose that another version of me, call him 'Twinny', on some other very similar world is likewise in a room that is slowly filling with carbon monoxide. Twinny doesn't have the hallucination. However, his friend sees the carbon monoxide levels of the room on a monitor (which Twinny had no access to), and, accordingly, bursts into the room to warn him. Twinny leaves the room, just as I did.

Now, what Twinny and I experience is subjectively indistinguishable (*ex hypothesi*), and our perceptual experiences justify each of our beliefs (we both know that we aren't prone to hallucinations etc.). However, only Twinny had an experience that can afford the opportunity for knowledge.[13] And that is the upshot of the fact that the justification that I had for believing that the room was full of carbon monoxide was caused by the fact that the room was full of

[13] Of course I could do the tests, gather the evidence and so forth (assuming I lived that long) and then I would be in a position to know that the room was full of carbon monoxide – but there and then, in circumstances as they were originally described, I would not be in a position to know.

carbon monoxide by a 'deviant' causal chain – whereas the chain in Twinny's case was 'non-deviant'; that is, it is only in Twinny's case that his experience is *in a suitable way* the upshot of the fact.

Generalising: we can say that the fact that *p*'s having caused a justification for the belief that *p* does not ensure that that justification is one that affords an opportunity for knowledge. So, I suggest, there must be a non-causal relation (as well as, perhaps, a causal relation) between the fact that *p* and a justification for believing that *p* if that justification is to afford an opportunity for knowing that *p*.

What is that non-causal relation? My answer is presumably clear: it is the same *mystery relation* as relates justified beliefs to the facts they are based on, and knowledge to that which is known. In other words:

(M3) For any propositions, *j* and *p*, if *j* affords the opportunity for knowledge that *p* then *j* is *mysteriously related* to the fact that *p*.

## 4 The mystery relation and acting for a reason

Finally, I want to convince you that the same mystery relation distinguishes actions done for a reason from mere bodily movements.

Consider the following example:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally. (Davidson 2001b, 79)

In this example the fact that the climber believes that loosening his grip would rid himself of danger (partially) explains why it is *pro tanto* rational for the climber to loosen his grip, and that belief also causes him to loosen his grip. However, he does not loosen his grip intentionally. This example thus (famously) creates a problem for a purely causal analysis of what it is to act intentionally: although there is a causal chain between the belief and the action, the agent does not act intentionally because the causal chain is 'deviant'.

Suppose that, had the climber acted on his beliefs and desires in the 'non-deviant' way, he would have acted intentionally. We can then say that Davidson's climber was *in a position* to do what he did intentionally, even though he didn't.[14] So what differentiates someone who *φs*

[14] Perhaps being in a position to do something intentionally is just having an intention to do it. I leave it for the reader to decide.

intentionally from someone (like the climber) who is in a position to *φ* intentionally, and, indeed, *φs* but does not do so intentionally?

The traditional 'answer' to this question is to say that if an agent acts intentionally then, *inter alia*, a belief that explains why their action is rational must cause them to do it *in the right way*.[15] So, again, assuming that this elusive 'right way' cannot be analysed in purely causal terms, we can say that if an agent acts intentionally then there is a non-causal relation (as well as, perhaps, a causal relation) between features of their psychology that explain why their action is rational and their action.[16]

And again, owing to the analogousness of this case to the others already considered, I suggest that this non-causal relation is the very same *mystery relation* that was required of justified belief, of knowledge and of opportunity for knowledge-affording justifications. Thus:

(M4) If *A φs* intentionally then, for some proposition, *p*, the fact that *A* believed that *p* explains why it was *pro tanto* rational for *A* to *φ* and is *mysteriously related* to the fact that *A φ'*d.[17]

## 5 A summary of the examples

I have suggested that what distinguishes deviant causal chains from non-deviant ones in the contexts considered is that, in the non-deviant cases, the relata are not *merely* causally related. In particular, I have argued, the same non-causal 'mystery' relation is present in each case, so that:

(M1) For any proposition, *p*, if *A* has a justified belief that *p* then the fact that *A* believes that *p* is *mysteriously related* to some justification for it.

(M2) For any proposition, *p,* if *A* knows that *p* then the fact that *A* believes that *p* is *mysteriously related* to the fact that *p*.

(M3) For any propositions, *j* and *p*, if *j* affords the opportunity for knowledge that *p* then *j* is *mysteriously related* to the fact that *p*.

(M4) If *A φs* intentionally then, for some proposition, *p*, the fact that *A* believed that *p* explains why it was *pro tanto* rational for *A* to *φ* and is *mysteriously related* to the fact that *A φ'*d.

As an aside: we should note that while mystery relations are non-causal relations, they do not necessarily exclude causal relations. That is, for instance, it is quite possible that a justified

[15] Cf. 'An action is performed with a certain intention if it is caused in the right way by attitudes and beliefs that rationalize it.' (Davidson 2001c, 87)

[16] Note that this is a necessary but not sufficient condition.

[17] This formulation excludes, for convenience, instances where an agent does something for its own sake (e.g. I sang because I felt like singing). In adopting this formulation I am not claiming that such acts are not done intentionally, it is just more convenient to use this formulation.

belief may be both causally and mysteriously related to the justification it is based on. Indeed, depending on your views in these areas it might be that a causal relation is a necessary (but not sufficient) condition for the presence of a mystery relation.

In the following sections I will argue further that the mystery relation is a transitive, explanatory relation.

# 6 Mystery relations are explanatory relations

Now I want to convince you that mystery relations are explanatory relations. They may be other things also, but I aim to convince you that they are definitely explanatory. I will argue that in each of the cases of the mystery relation above there is a non-causal explanation of the *explanandum* that is lacking when the mystery relation is absent.

## 6.1 Explaining justified beliefs

It is generally acknowledged that the justification on which an agent's belief is based explains why they believed it (e.g. Harman 1970). What sort of explanation does it provide us with?

If we accept that the basing relation is partly a causal relation then, even though a causal analysis is insufficient for establishing it, it's still possible that the explanatory import of the basing relation is merely causal. So, is the way that a justified belief is explained by the justification on which it is based *merely* causal?

The fact that Sally heard a bear-like sound headed her way (in a forest that she knew to contain bears) is a justification for her to believe that a bear is chasing her. Moreover, her belief that a bear is chasing her is *based* on that justification for it. So we can say, as we have noted, that Sally believes that a bear is chasing her partly *because* she heard a bear-like sound. That is: a justification for Sally to believe something partially[18] explains why she believed it.

Now notice that for Eva a justification for her to believe something also explains why she believed it: she believes that her husband has been unfaithful *because* it appeared to her as though he was kissing another woman (which, indeed, he was). That is, for both Eva and Sally a justification for their belief explains their belief. However, the sense of the explanation provided by the justification in Sally's case seems to be importantly different to the sense it provides in Eva's case – and that difference cannot be characterised in causal terms (because a causal relation obtains in both cases).

[18] Other parts of the full explanation of her belief include, for instance, the fact that she knew the wood to contain bears.

So, there is a seemingly non-causal explanation of Sally's belief that is lacking in Eva's case, which is to say that the explanatory import of the basing relation is not merely causal (note that Eva's belief is not based on *anything*, so there is no similar non-causal explanation of it).

## 6.2 Explaining knowledgeable beliefs

Another example: suppose that Eva's husband is kissing Sean's wife. Sean also sees it happen, but straightforwardly concludes that his wife has been unfaithful to him. Sean knows that his wife has been unfaithful and he believes that his wife has been unfaithful partly *because* she has been unfaithful.

Now, it is also true of Eva that she believes that her husband has been unfaithful *because* he has been unfaithful, even though she doesn't know that he has – but, again like the justified belief case, the sense of the 'because' seems different. Even given that there is a causal explanation in Sean's case, the fact that his wife is cheating on him partially explains the fact that he believes she is in a way that is *not merely* causal, because it explains it in a way that is qualitatively different to the *merely* causal explanation (i.e. the Eva case).

## 6.3 Mysterious relations are partial explanatory relations

I will spare the reader a rehearsal of this reasoning for the other cases considered and cut to the chase: wherever we compare a deviant causal case with a non-deviant case (in the contexts considered), it seems as though there is a non-causal explanatory relation in the non-deviant case that is lacking in the deviant one. Indeed, it seems to me that it is precisely the different character of the explanatory relations involved in the non-deviant case that allows us, in these contexts, to distinguish the non-deviant examples from the deviant ones.

So, since there appears to be a non-causal explanatory relation wherever we have a mystery relation, and since mystery relations, as we have established, are non-causal, I suggest that mystery relations are non-causal (partial) explanatory relations.

Thus, what separates Eva from Sean is the fact that the justification Sean has for his belief non-causally (as well as, perhaps, causally) explains his belief, whereas, for Eva it merely causally explains it.

Likewise, what separates me in my carbon monoxide filled room from Twinny is that the fact that the room was filled with carbon monoxide non-causally (as well as, perhaps, causally) explains why it appeared to Twinny as though his friend was warning him about the carbon monoxide, whereas it only causally explains why it appeared to me that way. And so on.

# 7 Mystery relations are transitive

Finally, I want to convince you that mystery relations are transitive. I will show that the conditions for transitivity are satisfied in all of the above examples, and I will argue further that the transitivity of the explanatory relations involved is the best explanation of why knowledgeable beliefs are mysteriously related to the believed facts.

## 7.1 It is transitive in the examples

When I congratulated my friend I knew that she had won an award, so the fact that I believed that she had won an award was mysteriously related to the fact that she had won an award. Moreover, I suggest, since I congratulated her for a reason, the fact that I congratulated her is mysteriously related to the fact that I believed that she had won an award. Can we conclude that the fact that I congratulated her is mysteriously related to the fact that she won an award? I think we can.

Consider: when I say that I congratulated my friend because she won an award it has the right sort of explanatory character, the fact explains my action *in the right way*. Likewise, when I say that Twinny left the room because it was full of carbon monoxide that too has the right sort of explanatory character. And that is, I suggest because the fact that the room is full of carbon monoxide is mysteriously related to his belief, which, in turn, is mysteriously related to his action. That is, I suggest that we can conclude that our respective normative reasons to act are mysteriously related to our respective actions because there is a chain of mystery relations connecting the normative reason to the action *and* the mystery relation is transitive.

We can see that these cases have the right sort of explanatory character by comparing them with a case in which the normative reason fails to be mysteriously related to my action: me in my room full of carbon monoxide. In this case when we say that I left the room because it was full of carbon monoxide that does not have the right sort of explanatory character. Of course there is perhaps a *sense* in which it is true (seemingly a strictly causal sense), but that is not the sense in which the expression would be conventionally understood. What marks out the sense in which the expression would be conventionally understood from this one is, as I have suggested, the presence of this non-causal explanatory relation; the *mystery* relation.

Why is the normative reason not related to my action in the carbon monoxide case but it is to Twinny's action? Because, I suggest, there is not a chain of mystery relations connecting my action to the fact that the room is full of carbon monoxide, while there is a chain of mystery relations connecting Twinny's action to that fact.

Moreover, while I will spare the reader a demonstration, I suggest that the same reasoning can be applied equally to all the other cases.

For clarity we can summarise this as follows. Firstly, some notation: for each case, let '*f*' stand for the believed proposition; '*j*' stand for the justification that the agent has for believing it; '*b*' stand for the fact that they believed it; and '*a*' stand for what the agent did. Table XIV-1 sets out the referents of these symbols for each case.

| **Example** | ***f*** | ***j*** | ***b*** | ***a*** |
|---|---|---|---|---|
| Award case | My friend won an award | I read that she had won an award | I believed that she had won an award | I congratulated her |
| CO[19] case | The room was full of carbon monoxide | It appeared to me as though my friend was warning me | I believed that the room was full of carbon monoxide | I left the room |
| Eva case | Her husband was unfaithful to her | It appeared to her as though her husband was kissing another woman | She believed that her husband was unfaithful | She sued for divorce |
| Climber case | Loosening his grip would rid him of danger | It appeared to him as though if he let go he would be freed from danger[20] | He believed that loosening his grip would rid him of danger | He loosened his grip |

**Table XIV-1:** The component facts in each example

My suggestion is this: in any circumstance in which a chain of transitive mystery relations would imply that the ends of the chain are mysteriously related, running through the reasoning just set out for each of the examples finds that the ends of the chain are, indeed so related. Letting '⇝' stand for the mystery relation, Table XIV-2 sets this out.

| **Examples** | **Observed relations** | | | **Should these obtain:** | | | **Do they obtain when they should?** |
|---|---|---|---|---|---|---|---|
| | $f \rightsquigarrow j$ | $j \rightsquigarrow b$ | $b \rightsquigarrow a$ | $f \rightsquigarrow b?$ | $j \rightsquigarrow a?$ | $f \rightsquigarrow a?$ | |
| Award case | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| CO case | × | ✓ | ✓ | × | ✓ | × | ✓ |
| Eva case | ✓ | × | ✓ | × | × | × | ✓ |
| Climber case | ✓ | ✓ | × | ✓ | ✓ | × | ✓ |

**Table XIV-2:** The transitivity of the mystery relation

[19] CO = carbon monoxide.

[20] You could supplant something more interesting here if you liked – this is just meant to be indicative.

Of course, the fact that the mystery relation happens to be transitive in these examples does not amount to proof that it is always transitive, but is at least evidence for that claim. In the next section I provide a different defence of the claim that mystery relations are transitive.

## 7.2 Knowledge and knowledge affording justification

My second argument for the transitivity of mystery relations is that it provides the best account of why it is that knowledgeable beliefs are mysteriously related to the believed facts.

To start with, consider that, to the extent that knowledge entails justified belief,[21] a knowledgeable belief must be *based* on a justification for it. That being so, it strikes me that if a belief is knowledgeable it must be *based* on (and therefore *mysteriously* related to) a justification that actually affords the opportunity for knowledge.[22] If these two claims are true, then, for any *p*, if *A* knows that *p*, the fact that *A* believes that *p* is based on some justification that affords the opportunity for *A* to know that *p*.

Summarising: for any *p*, a knowledgeable belief that *p*, being based on a justification that affords the opportunity for knowledge that *p*, must therefore be mysteriously related to a justification for the belief that *p* that is, in turn, mysteriously related to the fact that *p*. That is, to use the notation of the previous section, if the agent knows that *f* then: *f* is mysteriously related to *j*, *j* is mysteriously related to *b*, and *f* is mysteriously related to *b*.

So, if an agent knows that *p* then transitivity is true of the mystery relations between the belief that *p*, a justification on which it is based and the fact that *p*.

---

[21] Which, picking my battles, I will take as a given.

[22] Lehrer's gypsy lawyer case is a counterexample to this claim (or it would be if it were true):

> A lawyer is defending a man accused of committing eight hideous murders… There is conclusive evidence that the lawyer's client is guilty of the first seven murders. Everyone, including the lawyer, is convinced that the man in question has committed all eight crimes, though the man himself says he is innocent of all. However, the lawyer is a gypsy with absolute faith in the cards. One evening he consults the cards about his case, and the cards tell him that his client is innocent of the eighth murder. He checks again, and the cards give the same answer. He becomes convinced that his client is innocent of one of the eight murders. As a result he studies the evidence with a different perspective as well as greater care, and he finds a very complicated though completely valid line of reasoning from the evidence to the conclusion that his client did not commit the eighth murder… This reasoning gives the lawyer knowledge. Though the reasoning does not increase his conviction – he was already completely convinced by the cards – it does give him knowledge. (Lehrer 1971, 311–12)

I share Goldman's intuition that, 'To the extent that I clearly imagine that the lawyer fixes his belief solely as a result of the cards, it seems intuitively wrong to say that he knows—or has a justified belief—that his client is innocent.' (2012, 36 n8) The lawyer is in a position to know it, and depending on whether or not one thinks the tarot cards count as a justification, one might even say he has a justified belief – but I find it strange to say that he knows it if the presence of the only justification that affords the opportunity for knowledge is neither here nor there with respect to his actually having the belief.

Do we have reason to think that the mystery relation is transitive wherever it appears? Well we don't have any reason to doubt it, but here is another reason to think that it might be: the transitivity of the mystery explanation would explain *why* the belief that *p* is mysteriously related to the fact that *p* when the agent knows that *p*.

Although our intuition seems to suggest that a belief that *p* is non-causally (if also causally) explained by the fact that *p* if an agent knows it, it's not clear to me how the fact that *p* could *directly* explain an agent's belief that *p* in much the same way as it is was not clear to me how the fact that *p* could *directly* explain an agent's action.[23] However, given that our perceptual experiences intermediate the explanatory relations between the world and beliefs the transitivity of the relation that links them would ensure the connection that our intuitions suggest between the belief and the fact itself. So, I suggest, the best explanation of why knowledgeable beliefs are mysteriously related to the believed facts is that there is a transitive chain of mysterious relations that links the belief to the facts via the justifications.[24]

## 7.3 On whom is the burden of proof?

There are perhaps numerous points in these arguments to which one could object – but all they would do is undermine the *arguments* for the claim that mystery relations are transitive, they would not undermine the claim itself. Which leads me to my final point, in defence of the assumption that mystery relations are transitive: since I do not take it to be clear that the burden of proof is solely on me to prove that they are, it would need a further argument still to demonstrate that they aren't. In particular, since my view is that our working assumption of any particular sort of explanatory relation should be that it is transitive until proven otherwise, I take it that the burden of proof is actually upon those who would deny that the mystery relation is transitive.

[23] See my criticism of direct theories of normative reason explanation in § (X).

[24] It's worth noting that McDowell would reject this account. He notes that in the knowledge case 'appearances are no longer conceived as in general intervening between the experiencing subject and the world.' (1982, 472) So for McDowell it's not true that the justification intermediates between the belief and the fact in the knowledge case, rather the fact's having 'made itself perceptually manifest' is enough to do the job on its own.

# 8 Conclusion

I have for argued the following claims:

(M1) For any proposition, *p*, if *A* has a justified belief that *p* then the fact that *A* believes that *p* is *mysteriously related* to some justification for it.

(M2) For any proposition, *p,* if *A* knows that *p* then the fact that *A* believes that *p* is *mysteriously related* to the fact that *p*.

(M3) For any propositions, *j* and *p*, if *j* affords the opportunity for knowledge that *p* then *j* is *mysteriously related* to the fact that *p*.

(M4) If *A φs* intentionally then, for some proposition, *p*, the fact that *A* believed that *p* explains why it was *pro tanto* rational for *A* to *φ* and is *mysteriously related* to the fact that *A φ'*d.

I have argued that the mysterious relation in each case is *the same* relation, and that this relation is a non-causal, transitive, explanatory relation.

I should note: this is not a solution to the problem of deviant causal chains for any of these cases. I take the interesting questions for each case to be what this explanatory relation is, *why* it obtains, and when it obtains: I have called this the *mystery* relation precisely because the answers to those questions, *the interesting questions*, remains shrouded in mystery. All that I take myself to have done here is to offer some, hopefully, bland observations about this relation. Nonetheless, I hope that even these bland observations will be sufficient to make my case.

In the next chapter I will argue that this relation is transitive with the explanatory relation between, for instance, the fact that I believed that my friend had won an award and the fact that it was *pro tanto* rational for me to congratulate her.

# (XV)

# Mystery relations and why it is rational to act

*In which I say that mystery relations are transitive with the explanatory relation involved in explaining why it is rational. I label the sort of explanatory relation that obtains between (i) the fact that I believe that it is raining and (ii) the fact that it is* pro tanto *rational for me to take an umbrella, the 'E'-relation'. I argue that the mystery relation is transitive with the E'-relation. I show how this accords with our intuitions in some of the examples already considered.*

I want to say that the fact that my friend won an award explains why it was rational for me to congratulate her. I also want to say that the fact that the trees rustled does not explain why it was rational for Sally to run. However, in both cases, there is a chain of explanatory relations connecting the two facts.[1] So, if I am to say what I want to say, as noted in § (XIII)4, I need a principled reason for saying that all the explanatory relations involved in the first case are transitive, while the explanatory relations involved in the second case are not all transitive. In the last chapter I introduced the mystery relation with a view to making this case.

I have already suggested that the merely causal explanatory relation between the rustling of the trees and the fact that Sally heard a bear-like sound is not transitive with the explanatory relation between the fact that she believes that a bear is chasing her and the fact that it is *pro tanto* rational for her to run.

My argument in this chapter is this: the mystery relations between the fact that my friend had won an award and the fact that I believed that she had *are* transitive with the explanatory relation between the fact I believed that she had won an award and the fact that it is *pro tanto* rational for me to congratulate her. I argue that something that is not a feature of an agent's psychology (i.e. a belief or desire or what have you) explains why it is rational for them to do some action only if it is mysteriously related to a feature of their psychology that, in turn, explains why it is rational for them to do that action.

I put forward three arguments for the claim that mystery relations are transitive with the sort of explanatory relation involved in 'explaining why it is rational': firstly, I argue that they share many properties in common, and that the best explanation of *why* they share so many

[1] That is, the fact that my friend won an award explains why I read that she had won an award, which explains why I believed that she had won an award, which explains why it was *pro tanto* rational to congratulate her. Likewise: the fact that the trees rustled explains why Sally heard a bear-like nowise, which explains why she believed that a bear was chasing her, which explains why running was *pro tanto* rational. See § (XIII)4 for related diagrams.

properties is that they are the same sort of transitive, explanatory relation. Secondly, I argue that the best analysis of this sort of non-causal relation, *grounding*, takes it to be of a singular, transitive sort. Thirdly, I argue that this is the best account of why we might say, in ordinary language, that, for instance, it is rational for me to take my umbrella because it is raining.

# 1 Explaining why it is rational to act

## 1.1 The E'-relation

The fact that I believe that it is raining explains why it is *pro tanto* rational for me to take my umbrella. The fact that Sally believes that a bear is chasing her explains why it is *pro tanto* rational for her to run. Each of these cases involves a particular sort of explanatory relation – for expositional convenience it will help if we name it. Let us say the following:

> *Definition* For any proposition *p*, *p* is E'-related to the fact that it is *pro tanto* rational for *A* to *φ* if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*.

A clarification: the E'-relation is *not* a relation between, say, the fact that I believe that it is raining and the *act* of taking my umbrella. The E'-relation relates the fact that I believe that it is raining to *the fact that it is* pro tanto *rational for me to take my umbrella*. The E'-relation is just a particular sort of explanatory relation; it is whatever sort of explanatory relation it is that exists between those facts.

## 1.2 What are the properties of the E'-relation?

What can we say about the E'-relation? First, as we have already observed (see § (XIII)3.2.2), it is a non-causal explanatory relation. The fact that Sally believes that a bear is chasing her does not *cause* it to be *pro tanto* rational for her to run, and you should accept that even if you think it causes her to run. Causation just seems to be the wrong way to characterise this sort of relation. So, the E'-relation is non-causal.

Second, it is a transitive sort of explanatory relation. It seems right to say that if it's *pro tanto* rational for me to get some exercise, then that fact (partly[2]) explains why it is *pro tanto* rational for me to go swimming. Now suppose that it is *pro tanto* rational for me to get some exercise (partly) because I believe that exercise will lift my spirits. In such circumstances it also seems right to say that the fact that I believe that exercise will lift my spirits (partly) explains why it is *pro tanto* rational for me to go swimming.

[2] Together with, perhaps, the fact that I believe that going swimming is getting exercise, etc.

So what? Well, the fact that I believe that exercise will lift my spirits, *p*, explains why it is *pro tanto* rational for me to get some exercise, *q*. And the fact that it is *pro tanto* rational for me to get some exercise (i.e. *q*) explains why it is *pro tanto* rational for me to swim, *r*. *And* the fact that I believe that exercise will lift my spirits (i.e. *p*) also explains why it is *pro tanto* rational for me to swim (i.e. *r*). Thus: *p* explains *q*, *q* explains *r* and *p* explains *r*; so, at least in this instance, the explanatory relations involved satisfy transitivity.

Now, I submit, such explanatory relations must always be transitive: if it is rational for *A* to *φ* and *A* believes that *ψ*ing is a means to *φ*ing, then the fact that it is rational for *A* to *φ* will explain why it is rational for *A* to *ψ*. And whatever explains why it is rational for *A* to *φ*, together with *A*'s belief that *ψ*ing is a means to *φ*ing, will likewise explain why it is rational for *A* to *ψ*. So, the explanatory relation involved in explaining why it is rational to do something, i.e. the E'-relation, is a transitive relation.

Third, to the extent that ontological priority is a meaningful concept, the truth-makers of that which explains why it is *pro tanto* rational to do some action are ontologically *prior* to the rationality of actions. Consider: it's possible that someone's beliefs and desires/evaluative judgements (delete or replace as appropriate) may never align in a way that is sufficient to make any action rational (perhaps they have very odd desires, or normal desires, but weird beliefs, or normal desires and normal beliefs but just live in a dreadfully limited world) – the fact that no action is rational does not impinge on their ability to have beliefs and desires. So you can have beliefs and desires without there being any rational actions. However, the property of being rational cannot be instantiated without beliefs and desires. So, I suggest, the latter are ontologically prior to the former; which is to say that the E'-relation is (underpinned by) a relation of ontological priority.

Fourth, and relatedly, this sort of explanatory relation entails an 'in virtue of' claim. That is, if the fact that *p* explains why it is rational for *A* to *φ*, then we are seemingly always able to say that it is rational for *A* to *φ in virtue of* the fact that *p*. For instance, it is rational for Sally to run *in virtue of* the fact that she believed that a bear is chasing her. So, E'-relations entail 'in virtue of' claims.

## 2 A relation in common

I want to convince you that the mystery relation (a non-causal explanatory relation) is transitive with the E'-relation (also a non-causal explanatory relation). That is, I will argue that:

(M5) For any propositions, *p*, *q* and *r*, if *p* is mysteriously related to *q* and *q* is E'-related to *r* then *p* is E'-related to *r*.

Why should you believe this? Because, I argue, these relations are both the same sort of transitive, non-causal explanatory relations. I put forward three arguments for this claim: firstly, the pervasive similarities between the mystery relation and the E'-relation are best explained by the presence of a common relation. Secondly, the best account of explanatory relations of this sort, *grounding*, takes them to be (i) unitary (at least until proven otherwise) and (ii) transitive. Thirdly, accepting (M5) gives us the best explanation of why, in ordinary language, we might say, for instance, that it is rational for me to take my umbrella because it is raining.

A point of clarification: I am only arguing that these explanatory relations are of the same, *sort* (that is, that they belong to the same family of (transitive) explanatory relations), because that is enough for my argument for (M5). It is consistent with this claim to suppose that they are in fact the *same* explanatory relation,[3] however, it is likewise consistent with this claim to suppose that they aren't. I take no particular stance on whether or not they are the same explanatory relation just because I don't need to for my argument.

## 2.1 These relations are similar because they are of a common kind

### 2.1.1 There are a host of similarities between the relations

What are the similarities between the mystery relation and the E'-relation? We have already noted that they are both non-causal explanatory relations. Associated with their both being explanatory relations comes their both being asymmetric, irreflexive and non-monotonic. What else?

First, they are both transitive relations. In the previous chapter I demonstrated that the mystery relation is transitive. In the previous section I demonstrated that the E'-relation is transitive.

Second, they are both relations of ontological priority. I have already noted that that which explains why an action is rational must be ontologically prior to the rationality of the action. What of the mystery relation? The examples in the previous chapter relate, variously: facts about the world (or the agent's perception of it) to facts about what the agent believes; facts about the world to justifications for belief; and beliefs to actions. Presumably anyone who is not an idealist and finds some meaning in the notion of ontological priority will agree that facts

[3] And that might even be more parsimonious than thinking that they are different explanatory relations that belong to a single family.

of the former kind are ontologically prior to the latter.[4] That is, presumably: the world is ontologically prior to perceptual experiences of it, perceptual experiences are ontologically prior to beliefs, and beliefs are ontologically prior to actions (I have already argued for the last claim in the previous section). These don't all need to be true to make the case, but I think that they are.

Third, they both entail 'in virtue of' claims. Again, I have already noted that the E'-relation relation always entails an 'in virtue of claim', and so too does the mystery relation. Consider: it seems right to say that Sean believes that his wife has been unfaithful to him *in virtue of* the fact that she has been unfaithful to him and that Sally believes that a bear is chasing her *in virtue of* the fact that she heard a bear-like sound – in contrast it does not seem right to say that Eva believes that her husband has been unfaithful to her *in virtue of* the fact that she saw him kissing another woman.[5]

### 2.1.2 They are similar because they have some explanatory relation in common

I think that the best explanation of these similarities is that the two relations both are the *same* sort of transitive, non-causal, explanatory relation.

You might object to this. Perhaps you are sceptical of talk of ontological priority[6] or 'in virtue of'[7] relations. In which case you will doubt that there is much in the way of similarity that needs explaining. So be it. I, like many others,[8] increasingly take these to be meaningful concepts, and I suggest that the fact that the same form of words can be used in different cases is, at least, a *prima facie* reason for thinking that there is a common relation at work. Since there is a *prima facie* reason for thinking that there is a common relation I, following Audi, 'take the burden of proof to be on those who think there are different relations at work to show why.' (P. Audi 2012b, 689)

Another objection: perhaps you say that these similarities, at best, characterise a genus of non-causal explanation of which the mystery relation and the E'-relation are different species – their membership of the genus accounts for their similarity, but they are separated by their species-hood. That is, perhaps *all* of these similarities derive from there being an 'E-relation', where an E-relation is a non-causal explanatory relation that is not transitive with other

[4] And, I think, even some idealists can find a meaningful degree of ontological priority of suitably re-described facts of the former kind over facts about beliefs.
[5] Nor does it seem right to say that she believes that her husband has been unfaithful to her *in virtue of* the fact that he has; Eva's belief is neither knowledgeable nor justified.
[6] E.g. Hofweber (2009)
[7] E.g. 'We know we are in the realm of murky metaphysics by the presence of the weasel words "in virtue of ".' (Oliver 1996, 48)
[8] See e.g. Rosen (2010), Fine (2012) and Audi (2012a).

E-relations, even if it is transitive with itself. Nothing would then force us to conclude that mystery relations and the E'-rational relation are *the same* sort of E-relation, which is what is required for us to admit the truth of (M5).

Of course, this is possible, and, indeed, the different examples are not alike in all respects. However, what needs to be shown is not that there are differences between the cases given, but that what differentiates them is such that they cannot be the same sort of E-relation. And, again, I take the burden of proof here to be on those who think that they aren't.

## 2.2 The best account of such relations takes them to be of one kind

There is a readily available analysis of the ontological underpinnings of these explanatory relations according to which they are the same, transitive explanatory relation: namely, *grounding*.

### 2.2.1 What are grounding relations?

Grounding is the 'in vogue' relation in contemporary metaphysics. Here are a few characteristic grounding claims[9]:

> 1. Mental facts obtain in virtue of neurophysiological facts;
>
> 2. Dispositional properties are grounded in categorical properties;
>
> 3. Legal facts are grounded in non-legal, e.g. social, facts;
>
> 4. Morally wrong acts are wrong in virtue of non-moral facts;
>
> 5. Normative facts are grounded in natural facts.
>
> (Correia 2010, 251)

Assuming that grounding is a *bona fide* relation; here are some things that are taken to be essential to grounding[10]:

- It is an explanatory relation;[11]
- It is a not *merely* causal relation[12];
- It is transitive;

[9] Note: the usefulness of grounding doesn't hang on the truth of these claims but whether or not grounding can be used to characterise what the claims are claims *about*.
[10] See e.g. Rosen (2010), Fine (2012) and Audi (2012a).
[11] As an aside: We should note that one could question whether grounding relations are the non-causal explanatory relations themselves or the ontological determination relations that underpins them. That question is, however, largely orthogonal to our discussion.
[12] Though there is no requirement that grounding relations *exclude* causal relations.

– 'The fact that *p* is grounded in the fact that *q*' can be characterised by locutions like '*p* is the case in virtue of *q*'; and

– It is a relation of ontological priority.

We should note an important distinction between a *full* and a *partial* ground. Here is a typical characterisation:

> *A* is a partial ground for *C* if *A*, on its own or with some other truths, is a ground of *C*. (Fine 2012, 50)

Now, although full grounds are typically taken to necessitate that which they ground[13], merely partial grounds are not. For instance, the possibility of castling with one's kingside rook is partially grounded in the fact that no pieces obstruct the move but that fact alone does not guarantee that one can castle with one's kingside rook (for instance, the king may be in check).

And while full grounds may necessitate that which they ground, neither full nor partial grounds need be necessary for that which they ground. To take a standard example: the fact that this ball is scarlet fully grounds the fact that it is red, but the former is not necessary for the latter (the ball could be vermillion or ruby).[14]

### 2.2.2 These explanatory relations are grounding relations

Grounding theorists aim to explain why a variety of relations in seemingly different contexts exhibit the same properties by suggesting that they all share the common 'grounding' relation (and then providing an analysis of that relation). As Audi remarks:

> Such pervasive similarity among such diverse subject matters cries out for explanation. I propose that what accounts for the similarity is simply that there is a single relation at work in each case. (P. Audi 2012b, 689)

In § 2.1.1, I argued that the mystery relation and the E'-relation have a host of properties in common. Having set out the properties of grounding relations above, we can now see that the properties that these relations have in common just are the properties of grounding relations.

Like grounding relations, these explanatory relations are non-causal, transitive, explanatory relations; they can be characterised by locutions like 'in virtue of'; they involve claims of ontological priority; and they contribute to necessitating that to which they relate although they need not be necessary for it to obtain. Following Audi's logic, then, the best account of why these relations share those properties is because they, too, are grounding relations.

---

[13] For example, see Rosen (2010), Fine (2012) and Audi (2012b). There is some dissent from this view (for example, see Chudnoff (2011) and Leuenberger (2014)).

[14] The analogousness of the concepts of full and partial grounds with the concepts of full and partial explanation invoked throughout this discussion is presumably clear.

Moreover, the distinction between full and partial grounds readily accommodates what we earlier observed about the strictly partial contribution of the fact that Sally believes that a bear is chasing her to the explanation of why it is rational for her to run: we can say that the *partial* explanation *partially* grounds that which it explains. It similarly provides us with a ready characterisation of the strictly partial way in which the fact that it is raining explains why I believe that it is raining even when I know that it is raining (for it does not do so on its own): again we can say that the former fact *partially* grounds the latter.[15]

So here is another reason to believe (M5): there is a well-developed analysis of the sort of explanations that both the E'-relation and mystery relations appear to exhibit, *grounding*, which takes them to both be a common, non-causal, transitive, explanatory relation. That is, the best available account of the sort of explanations involved in the E'-rational and mystery relations entails the truth of (M5).

## 2.3 Their transitivity makes sense of ordinary language

A final consideration in support of (M5) is that, provided that you agree with (M1)-(M4), it provides the best account of ordinary language expressions in which non-psychological facts are said to explain why it is rational to do something.

Some things we might readily say: the fact that Sally heard a bear-like sound (in a wood that she knew to contain bears) at least partly explains why it was rational for her to run. The fact that my friend won an award at least partially explains why it was rational for me to congratulate her. It was rational for me to take my umbrella because it was raining.

As with normative reason explanations of action, there are three possible accounts of what is going on in these sorts of explanations: either the *purported explanans* explains the *explanandum* elliptically, directly or indirectly. Now, I suggest that the same arguments against the elliptical and direct accounts of normative reason explanation also apply to the elliptical and direct accounts of the explanation of why it is rational (see § (X)). In contrast, provided that you agree with my account of mystery relations, the claim that they are transitive with the E'-relation (i.e. (M5)) furnishes us with an account of these explanations that is thoroughly natural.

[15] Our readiness of to talk in terms of *grounds* for belief or *grounds* for action is, perhaps, further grist to this mill.

# 3 When non-psychological facts explain why it is rational

Recall the challenge set out in § (XIII)4: I need a principled account of why it is that, for instance, the explanatory relations from the believed fact to the rationality of the action are transitive in the 'award' case but not in the 'carbon monoxide' case. What is that account?

It is this: something that isn't a direct feature of an agent's psychology can explain why it is rational for them to do some action *only if* it is mysteriously related to a feature of their psychology that explains why that action is rational. In particular, a merely causal explanatory relation is not sufficient.

## 3.1 Some examples

It will help make the account clear if we revisit some of the different cases considered. The diagrams below set out three cases. I have labelled the explanatory relations as follows: instances of the mystery relation are marked '⇝'; instances of the E'-relation are marked 'E''; and *merely* causal relations are marked 'c'.

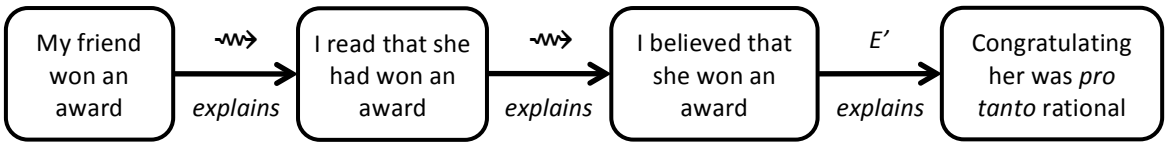


**Figure XV-1:** The explanatory relations in the *award* case.

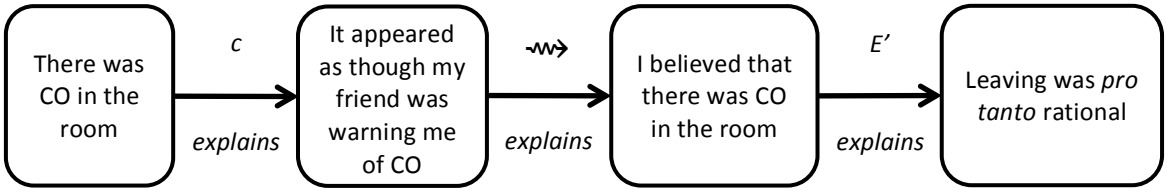


**Figure XV-2:** The explanatory relations in the *carbon monoxide* case

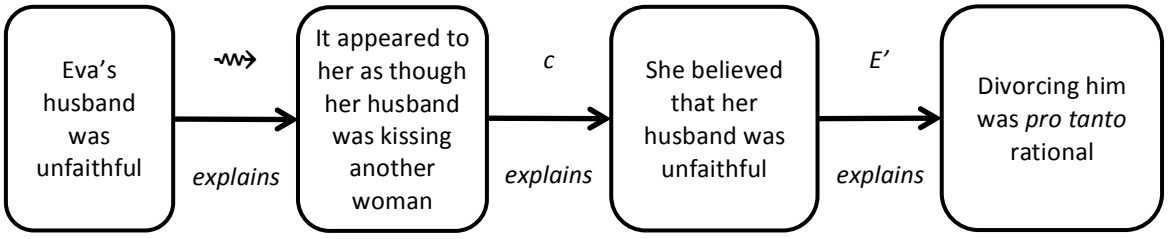


**Figure XV-3:** The explanatory relations in the *Eva* case

I will run through each of these examples in turn. In the award case (Figure XV-1) there is a chain of non-causal explanatory relations of a common sort that links the fact that my friend won an award to the fact that it was *pro tanto* rational for me to congratulate her. In particular: the fact that my friend won an award is mysteriously related to the fact that I read

that she had won an award in the newspaper, which is mysteriously related to the fact that I believed that she had won an award, which is E'-related to the fact that it was *pro tanto* rational for me to congratulate her.

Now, since the mystery relation is transitive, and transitive with the E'-relation (from (M5)), that means that the fact that my friend won an award is E'-related to the fact that it was *pro tanto* rational for me to congratulate her; that is: the fact that my friend won an award explains why it was *pro tanto* rational for me to congratulate her.

In contrast, in the carbon monoxide case, there is no chain of non-causal explanatory relations linking the fact that there was carbon monoxide in the room to the fact that it was *pro tanto* rational for me to leave it. The fact that there was carbon monoxide in the room is (at best) *merely* causally related to the fact that it appeared to me as though my friend was warning me about the carbon monoxide. However, the fact that it appeared to me as though my friend was warning me about the carbon monoxide *is* mysteriously related to the fact that I believed that there was (because I based my belief on the way things appeared to be), which is E'-related to the fact that it was *pro tanto* rational for me to leave. So, again, given the transitivity of these relations, the fact that it appeared to me as though there was carbon monoxide in the room *does* explain why it was *pro tanto* rational for me to leave. [16]

Finally, in Eva's case, there is likewise no chain of non-causal explanatory relations linking the fact that her husband was unfaithful to the fact that it was *pro tanto* rational for her to divorce him. While that fact *is* mysteriously related to the justification Eva has for believing it (i.e. that it appeared to her as though he was kissing another woman), since that justification is only *merely* causally related to her belief, the non-casual explanatory chain does not extend from the fact in the world, through the justification, to the rationality of the action. That is, neither the fact that Eva's husband was unfaithful to her, nor the fact that it appeared to her as though he was kissing another woman explains why it was *pro tanto* rational for her to divorce him.

[16] It will presumably also be clear how the same line of reasoning should lead us to the conclusion that the rustling of the trees does not explain why it is (*pro tanto*) rational for Sally to run. That is, the rustling of the trees is merely causally related to her hearing a bear-like sound, so the explanatory between the two facts is not transitive with the E'-relation, as it would need to be in order to explain why her action was (*pro tanto*) rational.

The table below summarises these cases:

| Example | | The believed fact/justification for believing it | Explains why it was *pro tanto* rational to act? |
|---|---|---|---|
| Eva case | ***f*** | Her husband was unfaithful to her | × |
| | ***j*** | It appeared as though her husband was kissing another woman | × |
| CO case | ***f*** | The room was full of carbon monoxide | × |
| | ***j*** | It appeared as though my friend was warning me | ✓ |
| Award case | ***f*** | My friend won an award | ✓ |
| | ***j*** | I read that she had won an award in the newspaper | ✓ |

**Table XV-1:** A summary of what explains and what doesn't in each case

# 4 Conclusion

I have argued that something that is not a direct feature of an agent's psychology can explain why their action is *pro tanto* rational only if it is mysteriously related to a feature of their psychology that, in turn, explains why their action is *pro tanto* rational. I have argued for this on the basis that the mystery relation is transitive with the E'-relation, while merely causal relations are not. I consider some objections to this proposal in the Appendix to this chapter. The next, and final, chapter combines the insights of this discussion to set out my theory of reasons.

# Appendix

## A.1 Objections

**Objection 1** The challenge was for you to show why it was that (i) the fact that your friend won an award explained why it was rational for you to congratulate her, despite (ii) the fact that the room was full of carbon monoxide didn't explain why it was rational for you to leave. Your 'answer' is that it is a mystery, but that's no sort of answer! You are just re-labelling the problem that was already diagnosed. What progress has really been made? Don't we want to know *why* this relationship is present in (i) and not in (ii)?

**Response** Well, here is some progress that has been made: we've identified that the mystery relation is part of some family of explanatory relations, where members of that family have certain properties. That is at least the starting point for a systematic investigation of why the relationship is present.

More generally, however, I am not trying to answer the question of why the relationship is present. The question I am trying to answer is why the explanatory chain in (i) is transitive and why it isn't in (ii). My answer is that *there is* a chain of (mysterious) explanatory relations in (i) that there isn't in (ii) *and* that these *mystery* relations are transitive with the E'-relation (i.e. (M5)). If you accept these two claims that is enough to answer the question that I am trying to answer.

The question as to *why* there is a mystery relation is present in (i) but isn't in (ii) is a deeper (more *distal*) and perhaps more interesting question than the one I am answering. But answering it is beyond the scope of what I need to do to make my case. In identifying the mystery relation as an explanatory relation, and showing that it is transitive with the E'-relation, and showing that it obtains in (i) but not in (ii), I have an answer to the challenge posed, and I don't need more than that.

**Objection 2** If grounding-based formulations of physicalism are true, then you are forced into the absurd conclusion that, say, the fact that Sally's brain is in state *B* explains why it is rational for her to run.

Much of your argument for (M5) was based on the claim that the best explanation of all the grounding-like properties that the mystery relation and the E'-relations share is that they both involve a common, explanatory relation. It was for that reason that you suggested that they were transitive, so, you suggest:

> (a) All explanatory relations that exhibit grounding-like properties are transitive with each other.

Grounding-based formulations of physicalism (e.g. Correia 2010; Kroedel and Schulz 2016) hold that all mental facts are *grounded* in physical ones. Now, grounding-based formulations of physicalism take the relation between mental and physical facts to exhibit exactly those properties that you said that the E'-relation and mystery relation exhibit.[17] Thus:

> (b) All mental facts stand in a particular explanatory relation to some physical facts and that explanatory relation exhibits grounding-like properties.

Thus, from (a) and (b), together with your claim that the E'-relation exhibits grounding-like properties, we infer that the explanatory relation between the mental and the physical is transitive with the E'-relation.

[17] For instance: The relation is a non-causal explanatory relation. It can be characterised by the 'in virtue of' locution. Etc.

So, some physical fact (e.g. the fact that Sally's brain is in state *B*) non-causally explains the fact she believes that it is rational for her to run. And since that is transitive with the E'-relation, we can conclude that the fact that Sally's brain is in state *B* explains why it is *pro tanto* rational for her to run. But that is absurd!

**Response** I agree that it is absurd to say that facts about Sally's brain state explain why it was *pro tanto* rational for her to run. So, to preserve my theory in the face of this conclusion I must reject either (a) or (b). [18]

As it is, I'm not convinced that (b) is true – in particular, it's not clear to me that facts about an agent's brain state *explain* their mental state. However, the claim that reduction relations are explanatory relations is true in a number of popular construals of physicalism beyond just grounding-based formulations[19] – so rejecting it is not without its costs.

Is there a way to render my theory consistent with such construals of physicalism, by rejecting (a), that doesn't also undermine my argument for (M5)? I suggest that there is.

Recall that in my argument for (M5), I considered the possibility that the E'-relation and the mystery relation involve different species of a genus of non-causal explanatory relations, *E-relations*, and that it is their belonging to that genus that accounts for the properties they have in common. The argument was that E'-relations and mystery relations need not involve the same relation to exhibit the same properties – being members of the same family is sufficient. My response to that possibility was Audi's: that the burden of proof is on those who take the relations to be different to show that they are different.[20]

With respect to mystery relations, I can think of no compelling reason as to why we should think that the sort of explanation involved is entirely different to the sort involved in the E'-relation. However, unlike the mystery relation, to the extent that the relation between the mental and the physical really is a non-causal explanatory relation (i.e. given that we assume (b)), I think that we have reason to think that it does not involve the same sort of non-causal explanatory relation as the E'-relation. That is, I think that in this case we do have a reason for thinking they involve different sorts of explanation, but in the case of the mystery relation, we don't.

What is the reason we have? I argue that the sort of explanatory relation that is meant to exist between the mental and the physical is not transitive with the sort of explanation the

[18] At least so long as we assume that, on any grounding-based version of physicalism, Sally's brain state is the 'ground' of her mental state.
[19] This is what Crane (2000) calls 'conceptual reduction' (as opposed to 'ontological reduction').
[20] Schaffer (2009, 377) makes similar remarks.

E'-relation involves – and that this is a reason for thinking that they do not involve the same relation.

Let us introduce a distinction (which is meant to be intuitive, but which will be substantiated with examples) between two sorts of non-causal explanation: *vertical* and *horizontal*.[21] Say that *vertical* explanatory relations, *$E^v$-relations*, are explanatory *reduction* relations (if there are such things), like those between the mental and the physical. Say that *horizontal* explanatory relations, *$E^h$-relations*, are like those involved in E'-relations and mystery relations. What I will need to demonstrate is that horizontal explanatory relations are not transitive with vertical explanatory relations (i.e. that (a) is false).

To start: consider that the fact that this slice of cake is the biggest explains why it would be impolite to take it. This, I submit, is a *horizontal* explanatory relation – the fact that the cake is biggest makes taking it impolite in the same way that a belief that it's cold explains why it is rational to turn on the heating.

Now given that reduction is an explanatory relation, as the 'conceptual reduction' physicalist supposes, facts about the microphysical properties of the slice (and the other slices) explain why it is the largest slice. I suggest that this explanatory relation, *qua* reduction relation, is the same sort of vertical explanatory relation, i.e. the $E^v$-relation, as the relation between mental and physical facts.

But, I suggest, it is odd to say that facts about the microphysical structure of the slice of cake are a part of the explanation of why it's impolite to take it. That is, we cannot infer from (i) the fact that facts about the microphysical structure of the cake (partly) explain why it is the largest slice; and (ii) the fact that the fact that it is the largest slice (partly) explains why it is impolite to take it; to (iii) the conclusion that facts about the microphysical structure of the cake (partly) explain why it is impolite to take it. In other words, the transitivity of explanation breaks down. Now, since the $E^h$-relation is a transitive relation, we have a reason to think that the $E^v$-relation is not transitive with the $E^h$-relation, so they are distinct sorts of explanatory relation.

This particular argument relies on the reduction of macro-physical to micro-physical facts being relevantly analogous to the reduction of mental facts to physical ones. However, I

[21] I have appropriated and re-purposed this terminology from Jaegwon Kim (2003). Kim talks in terms of 'vertical determination' and 'horizontal determination' – whilst vertical determination is close to what I characterise as 'vertical explanation', Kim's notion of horizontal determination is explicitly causal. For that reason I distinguish between vertical *explanation* and horizontal *explanation*.

cannot see how transitivity with $E^h$-relations could fail for the former, (as I have argued that it does) but succeed with the latter, so I have inferred that it doesn't.[22]

In contrast, as I have argued above, the transition from a mystery relation to an E'-relation seemingly does involve a transitive form of explanation: the claims we end up making if we take the relation to be transitive there (e.g. the fact that it is cold explains why it's rational for me to put the heating on) seem commonplace in comparison to the preposterousness of the claims we make if we assume that vertical explanations are continuous with horizontal ones. My point is just that while $E^v$-relations are 'transitive with' each other, they are not 'transitive with' $E^h$-relations.

Thus, I think that there is a credible way of rejecting (a) that does not undermine my argument for (M5), so that the proponent of grounding-based formulations of physicalism (and *conceptual reduction* more generally) can accept my theory without arriving at the absurd conclusion that Sally's brain state explains why it is rational for her to run. The only cost to such a physicalist is that they must admit that there are at least two kinds of grounding relation that, despite sharing a host of properties, are not transitive with one another. However, this is not a cost I need pay since I anyway don't think that reduction relations are explanatory (that is, my preferred response to this objection is to reject (b)).

**Objection 3** Deviant causal chains affect many other analyses of causation, for instance:

> It would not work to say that the heat of the oven cooks the chicken if and only if the heat of the oven causes the chicken to be in a cooked state. The heat of the oven might trigger some microwave activity elsewhere which causes the chicken to be in that state; in this case the heat would not have cooked the chicken. (Stout 2010, 161)

That being so, the mystery relation is presumably not only restricted to the cases you've considered. Supposing that the mystery relation differentiates deviant from non-deviant cases, we might say this: if the oven cooks the chicken then the fact that the oven cooked the chicken is *mysteriously* related to the fact that the chicken reached a cooked state.

So, let's imagine a twice deviant carbon monoxide case. Instead of causing you to hallucinate in the 'normal' way, the carbon monoxide causes a creature in the room to hallucinate that you are attacking it, and, unbeknownst to you, it injects you with hallucinatory venom. That then makes you hallucinate that your friend is warning you about the carbon monoxide and so on.

[22] Indeed, the fact that it is absurd to claim that the fact that Sally's brain is in state *B* explains why it is *pro tanto* rational for her to run is clear evidence that the explanatory relations involved are not transitive.

In the original example the carbon monoxide made you hallucinate in a non-deviant way. In this new case it made you hallucinate in a deviant way. But if what differentiates deviant cases from non-deviant ones is the mystery relation, then that must meant that, in the non-deviant case, the fact that there was carbon monoxide in the room *is* mysteriously related to the fact that it appeared to you as though your friend was warning you about the carbon monoxide. That being so, given the transitivity of the mystery relation, the fact that there was carbon monoxide in the room ought to explain why it was rational for you to leave it.

**Response** First, I never offered the mystery relation as an analysis of what differentiates deviant causal chains from non-deviant ones *in any context*. It was specifically restricted to the cases considered.

Second, I do not think it should be extended to other cases. What distinguishes deviant causal chains from non-deviant ones is *mysterious* in the mental case *because* it involves the mental (or, at least, the *representational*). In the case of cooking the chicken, while I don't have the solution to the problem, I don't think there is the same fundamentally mysterious problem at work. Whatever the mystery relation is, it is, I suggest, to do with the rational faculties of agents, whereas, it seems to me, the question of deviant causal chains in mere causal cases is a mere question of mechanism.

Perhaps you don't find this response very compelling. I'm afraid I don't have a more compelling one; should this prove to be an insuperable difficulty then I should have to look for some other analysis as to why facts about the world can explain why our actions are rational. However, I struggle to find this objection insuperable.

# (XVI)

# A new theory of reasons

*In which I set out my theory of reasons. I discuss what explanatory rationalism says about the application of each reason expression to the case where I take my umbrella having seen that it is raining. I show how explanatory rationalism solves the problems faced by other theories. I suggest that the best theory of reasons is a pluralist theory of reason that combines explanatory rationalism and favourism; I call this theory 'new pluralism'. I show how explanatory rationalism enables new pluralism to meet the main challenge to pluralist theories.*

In § (V), I argued that if we are to solve all of the problems discussed in §§ (II)-(VI) then we need a new family of claims about reasons. In § (VI), I suggested that explanatory rationalism was the new family of claims that we needed, however, I noted, The Explanatory Exclusion Problem presented a significant challenge for it. The intermediating chapters have argued that The Explanatory Exclusion Problem is not the problem it seems to be, and that it is therefore no obstacle to explanatory rationalism's solving the problems discussed in §§ (II)-(V).

Now it is time to discuss how explanatory rationalism solves these problems. The answer is perhaps obvious: explanatory rationalism solves these problems by rejecting the troublesome views that gave rise to them; that is, explanatory rationalism rejects favourism, psychologism and deliberativism. More generally, explanatory rationalism is not susceptible to similar sorts of problems because it is consistent both with the idea that agents always act for psychological reasons, and with the idea that they sometimes also act for normative reasons. Because of this, I argue, explanatory rationalism is the best univocal account of what it is to be a reason.

However, you will recall from § (V) that I think that our theory of reasons ought to be pluralist because I share the *two senses* intuition. To that end, I present *new pluralism*: I suggest that one sense of what it is to be a reason is *explanatory rationalist* and the other sense is *favourist*. That is my theory of reasons.

In what follows I revisit explanatory rationalism, and consider what it says about what my reasons were when I saw that it was raining, and consequently took my umbrella. I then show how explanatory rationalism solves the problems that affect other theories. Finally, I set out new pluralism and show how it addresses the main challenge to pluralist theories.

# 1 Explanatory Rationalism: Revisited

Recall what explanatory rationalism has to say about each reason expression:

| Reason expression | Explanatory rationalism |
|---|---|
| For any *p, p* is a reason for *A* to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. |
| For any *p, p* is a reason for *A's φing*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and *p* makes *A's φing*, in some respect, worth doing. |
| For any *p, p* is a reason *A* has to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. |
| For any *p, p* is *A's* reason for *φing*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and explains (in the right way) why *A φ'd*. |

**Table XVI-1:** Explanatory rationalism

It may help put this theory into context if we consider an example. Before doing so, however, there is something that perhaps still needs to be made explicit: what it means to explain *in the right way* why someone acted. For the purpose of this discussion I will assume that if some fact is mysteriously related to the fact that the agent did what they did, then it explains it in the right way. I discuss this assumption further in § A.2 of the Appendix to this chapter.[1]

That clarification having been made, let us re-consider the following example: I look out of the window and see rain. I know that it's raining, so I take my umbrella when I leave the house.

## 1.1 The explanatory relations involved

What are the explanatory relations in this example? Well, the fact that it was raining is mysteriously related to the fact that it appeared to me as though it was raining, which is mysteriously related to the fact that I believed that it was raining, which is mysteriously related to the fact that I took my umbrella. Furthermore, the fact that I believed that it was raining explains why it was *pro tanto* rational for me to take my umbrella. Diagrammatically:

[1] As an aside: it is worth also noting that this assumption does not preclude the possibility that a causal explanatory relation is also necessary for a reason to explain an action in the right way: in particular, it might be that (at least when it comes to action) a causal explanatory relation between a reason and an action is a necessary (but not sufficient) condition for a mystery relation. In which case explaining why someone did what they did *in the right way* involves *both* a causal explanatory relation and a mystery relation.

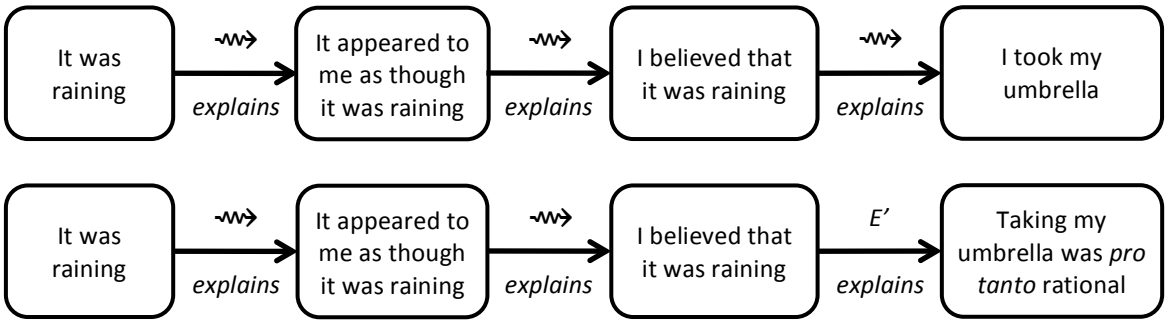


**Figure XVI-1:** The explanatory relations involved when it was raining

So, given the transitivity of the mystery relation with the E'-relation, the following facts all (partially) explain why it was *pro tanto* rational for me to take my umbrella: the fact that it was raining, the fact that it appeared to me as though it was raining and the fact that I believed that it was raining. And, given the transitivity of the mystery relation, these same facts are all also mysteriously related to the fact that I took my umbrella.

## 1.2 The reasons there were and the reasons I had to take my umbrella

According to explanatory rationalism, explaining why it is *pro tanto* rational for someone to do some action is both necessary and sufficient for being a reason for them to act *and* for being a reason that they have to act.[2] So, the fact that it was raining, the fact that it appeared to me as though it was raining and the fact that I believed that it was raining were all reasons for me to take my umbrella *and* reasons that I had to take my umbrella.

## 1.3 The reasons for taking my umbrella

According to explanatory rationalism, the conditions for being a reason for (or against) doing something are different from the conditions for being a reason to (or not to) do it. In particular, explanatory rationalism maintains that a reason for doing something must both be a reason to do it *and* make it, in some respect, worth doing.

The fact that it is raining is a reason for me to take my umbrella and it makes taking it, in some respect, worth doing, so it is a reason for my taking my umbrella. However, the fact that it appeared to me as though it was raining, and the fact that I believed that it was raining do not make taking my umbrella, in any respect, worth doing,[3] so, despite being reasons for me to take my umbrella, they aren't reasons for taking my umbrella.

[2] Note that since, according to explanatory rationalism, the conditions for being a reason there is to act are the same as being a reason that one has to act, explanatory rationalism takes these two expressions to be coextensive – that is, to pick out the same kind of reason. Thus explanatory rationalism entails (F8) (reasons agents have to act area reasons for them to act).

[3] Recall the discussion of this point in § (I)4.2 in particular.

If you are concerned that it sounds odd to say that something is a reason to do something but not a reason for doing it, please refer to the discussion of § A.3 of the Appendix to this chapter.

## 1.4 My reason for taking my umbrella

Finally, explanatory rationalism holds that the reason for which an agent acts is anything that both explains why it was *pro tanto* rational for them to act[4] and explains why they acted (in the right way), which is to say that any reason for an agent to act that explains why they acted (in the right way) is their reason for acting.[5]

I have assumed that the mystery relation is sufficient for explaining why someone acted in the right way. This means that any reason for some agent to do some action that is mysteriously related to the fact that they did it is amongst their reasons for doing it.

I have noted that the following facts were all both reasons for me to take my umbrella and are mysteriously related to the fact that I took it: the fact that it was raining, the fact that it appeared to me as though it was raining, and the fact that I believed that it was raining. Thus these were all reasons for which I took my umbrella.

I have said that the fact that it was raining was amongst my reasons for taking my umbrella. Consider also that the fact that it was raining was a normative reason for me to take my umbrella: so, I acted for a normative reason. Some theorists hold that one acts for a normative reason only if one knows it (as I did in this case); in § A.5 of the Appendix to this chapter I consider how explanatory rationalism could explain why that should be so.

## 1.5 A summary

Table XVI-2 provides a summary of what explanatory realism has to say about which reason expressions apply to which fact in this example.

[4] An implication of this formulation is that an agent could do something for a reason that was *pro tanto* but not all things considered rational for them to do. I discuss this implication further in § A.4 of the Appendix to this chapter.

[5] It is worth noting that the formulation of explanatory rationalism thus entails the *prima facie* reasonable claims (F7) and (F9) (the claims that connect acting for a reason with there being a reason to act) as well as (D2) (the claim that the agent's reason for acting always explains their action).

| The putative reasons | A reason for me to take my umbrella | A reason for my taking my umbrella | A reason I had to take my umbrella | A reason for which I took my umbrella |
|---|---|---|---|---|
| It was raining | ✓ | ✓ | ✓ | ✓ |
| It appeared to me as though it was raining | ✓ | × | ✓ | ✓ |
| I believed that it was raining | ✓ | × | ✓ | ✓ |

**Table XVI-2:** An application of explanatory rationalism

The facts considered above do not exhaust the reasons that there are in this example,[6] however they should hopefully be indicative of the sorts of claims that explanatory rationalism makes. For reference, I provide the same table for several of the examples considered in previous chapters in § A.1 of the Appendix to this chapter.

## 2 Solving the problems

Explanatory rationalism solves the problems considered in §§ (II)-(VI) by rejecting the problematic theses that gave rise to them. That is, explanatory rationalism rejects all of the following claims:

(FAV) For any *p, p* is a reason for *A* to *φ* only if *A*'s *φ*ing, is in some respect, worth doing.

(PSY) For any *p,* if *A φs* for the reason that *p* then *p* is a feature of *A's* psychology.

(DEL1) For any *p,* if *p* is a consideration in light of which *A φs* then *p* is *A's* reason for *φing.*

(DEL2) For any *p,* if *p* is *A's* reason for *φing* then *p* is a consideration in light of which *A φs.*

Many will find explanatory rationalism's rejection of these views unpalatable. Indeed, it will seem to many that at least some of these claims are themselves *prima facie* reasonable. I do not take that to be the case – rather I think that any resistance to rejecting these claims comes from one's theoretical commitment to them, and not from their inherent plausibility.

Showing that explanatory rationalism can solve the problems considered in §§ (II)-(VI) is, of course, not a demonstration that it is consistent with all of the *prima facie* reasonable claims set out in those chapters. Indeed, it is possible that there are other *prima facie* reasonable

[6] For instance, the fact that my umbrella would keep me dry is also a reason for me to take it (given that I know that it would), and a reason for which I took it.

claims out there that, when combined with those that I have considered, create new problems that are specific to explanatory rationalism.

Against that possibility I have two responses: firstly, it is for the sceptic to generate such claims, not for me to prove that none exist (how could I prove that?). Secondly, and perhaps more compellingly, explanatory rationalism is to some extent inoculated from the conventional form of these problems because it acknowledges both that agents always act for psychological reasons, and that agents also can, and often do, act for normative reasons. Because it can reconcile these seemingly competing theses, it is immune from problems that arise from denying either, and, for that reason, I suggest that it is the best available univocal theory of reasons.

# 3 New pluralism

Explanatory rationalism may be the best available univocal theory of reasons, however, I think that our theory of reasons ought to be pluralist. Recall that in § (V) I introduced the 'two senses' intuition: I suggested that there is a sense in which Sally ran for a reason and a sense in which she didn't; similarly, I suggested that if I believe that there is milk at home even though there isn't, then there is a sense in which I don't have a reason to buy milk and a sense in which I do. In light of this observation I suggest that we should not stop with the univocal account of reasons provided by explanatory rationalism on its own; instead, we ought to adopt a pluralist theory of reasons.

The sense in which Sally runs for a reason is the sense in which running intentionally, deliberately and purposefully is running for a reason. Likewise, the sense in which I have no reason to buy milk is the sense in which, if it is not rational for me to do something, I have no reason to do it. Over the previous chapters I have argued that the best way to characterise the sense of these expressions for which these claims are true is explanatory rationalism.

In contrast, the sense in which Sally does not run for a reason is the sense in which there is no reason for her to run because no good will come from her running; it is in no respect worth doing. Likewise, the sense in which I have a reason to buy milk is the sense in which one has a reason to do something because doing that thing *is*, in some respect, worth doing. Thus, the sense in which Sally does not run for a reason, and the sense in which I have a reason to buy milk is characterised by the *favourist* family of claims about reasons.

So, to the extent that one shares the 'two senses' intuition, I suggest that the best theory of reasons takes one sense of each reason expression to be explanatory rationalist, and the other sense to be favourist. This is the *new pluralist* theory of reasons.

| Reason expression | Sense *A* (Explanatory rationalism) | Sense *B* (Favourism) |
|---|---|---|
| For any *p*, *p* is a reason for *A* to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p*, *p* is a reason for *A's φ*ing… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and *p* makes *A*'s *φ*ing, in some respect, worth doing. | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p*, *p* is a reason *A* has to *φ*… | …if and only if *p* explains why it is *pro tanto* rational for *A* to *φ*. | …if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing. |
| For any *p*, *p* is a *A*'s reason for *φ*ing… | … if and only if *p* explains why it is *pro tanto* rational for *A* to *φ* and explains (in the right way) why *A φ*'d. | …if and only if *p* makes *A*'s *φ*ing, *all things considered*, worth doing and explains (in the right way) why *A φ*'d. |

**Table XVI-3:** New Pluralism

While I think that new pluralism is the best way to capture our myriad intuitions about reasons, I think that explanatory rationalism is the *de facto* sense of reason expressions in ordinary language. It seems to me to be more of a strain to insist that I have a reason to buy milk even if I believe that I have plenty, than to insist that Sally has a reason to run although no bear is chasing her. Indeed, I feel more inclined to qualify the former – I might say that *there is a sense in which* I have a reason to buy milk – and I am so inclined, I think, because the sense in which I have a reason to buy milk is not the conventional sense of what it is to have a reason (*mutatis mutandis* for Sally).[7]

# 4 The challenge for pluralism

Recall the following *prima facie* reasonable claim:

(S1) Whenever we give an agent's reason for acting, *whatever the sense of the expression used*, we explain their action in a way that makes them seem rational.

In § (V)7, I noted that (S1) creates a problem for pluralism. The problem was as follows: a theory of reasons that is pluralist with respect to the reasons for which an agent acts, takes the 'agent's reason for acting' expression to have two different senses. In particular, such a theory holds that the reason-relation for each sense (i.e. the relation between the action and the

[7] Of course, the great virtue of explanatory rationalism is that there are occasions on which something is a reason that I have to do something in *both* the favourist and the explanatory rationalist sense. For instance, when I know that it is raining, the fact that it is raining is a reason for me to take my umbrella in *both* the favourist and the explanatory rationalist sense.

reason in virtue of which it is a reason) is different. The challenge for pluralist theories is thus to explain how it is that, if the reason-relations are different for the different senses, it is nonetheless the case that when we give an agent's reason for acting, *whatever the sense of the expression used*, we explain their action in a way that makes them seem rational. That is, if the expression 'the agent's reason for acting' has two different senses, how is (S1) true?

New pluralism's answer to this question relies on the following observation: whenever an agent acts for a reason in the favourist sense, they act for the same reason in the explanatory rationalist sense. To see why this is so, recall what favourism about the reasons for which we act claims:

- *Favourism about the reasons for which we act*: For any *p*, *p* is a *A*'s reason for *φ*ing if and only if *p* makes *A*'s *φ*ing, *all things considered*, worth doing and explains (in the right way) why *A φ*'d.[8]

If something makes an agent's action *all things considered* worth doing then it also makes it *in some respect* worth doing, so it is a normative reason for them to do it. I have suggested, in previous chapters, that a normative reason to do some action explains what the agent did in the right way only if it is mysteriously related to a belief that, in turn, both explains why the agent's action is *pro tanto* rational and is mysteriously related to the fact that they (intentionally) did it. Now recall that a belief that explains both why it is *pro tanto* rational for some agent to do something and why they did it is the agent's reason for acting in the *explanatory rationalist* sense.

Thus, a normative reason, *p*, is the agent's reason for acting in the *favourist* sense only if it is mysteriously related to the fact that the agent believes that *p*, where the fact that the agent believes that *p* is the agent's reason for acting in the *explanatory rationalist* sense. But, as I have argued, if *p* is mysteriously related to the fact that the agent believes that *p*, and the fact that the agent believes that *p* is, in the explanatory rationalist sense, their reason for acting, then the fact that *p* is *also* their reason for acting in the explanatory rationalist sense. Thus, whenever an agent acts for a reason in the favourist sense, they act for the *same* reason in the explanatory rationalist sense.

[8] While this thesis is rarely, if ever, explicitly advocated by favourists, I take it to be at least implied by their views. For instance, many favourists hold that an agent acts for a normative reason only if they know it (see the discussions of both § (X)4 and § A.5 of the Appendix to this chapter for some related discussion of these accounts). Now, whether or not they accept my indirect theory of normative reason explanation, such favourists still want to hold that a normative reason must explain the agent's action *in the right way* if it is to explain their action – and that is all that is required for them to agree with my characterisation.

And when an agent acts for a reason in the explanatory rationalist sense, their reason for acting explains both why they acted and why it was *pro tanto* rational for them to act. So this is why giving an agent's reason for acting always explains their action in a way that makes them seem rational (or at least *pro tanto* rational): because their reason for acting always explains why they did what they did and why it was (at least *pro tanto*) rational for them to do it.

New pluralism thus meets the challenge that (S1) creates for all pluralist theories by insisting that whenever something is an agent's reason for acting, it is a reason for acting in the explanatory rationalist sense, but sometimes it is *also* the agent's reason for acting in the favourist sense. New pluralism thus provides a plausible account of the truth of (S1) whilst maintaining its pluralist credentials.

# 5 Conclusion

Explanatory rationalism provides us with a univocal account of what it is to be a practical reason that does not suffer the failings of most contemporary theories of reasons. The great virtue of explanatory rationalism is thus that it does not generally compel us to make claims about reasons that are strange, counterintuitive or *prima facie* paradoxical, unlike most contemporary theories of reasons. In this respect, it is, I suggest, superior to favourism, deliberativism, and psychologism.

Moreover, I have suggested that *new pluralism*, a theory that combines explanatory rationalism and favourism, is the best theory of reasons; it combines all the virtues of explanatory rationalism with a way to satisfy the 'two senses' intuition, and it happily meets the major challenge to pluralist theories.

# Appendix

## A.1 Some more examples

For reference, Table XVI-4 shows what explanatory rationalism has to say about which reason expressions apply to which facts/propositions in some of the examples considered in earlier chapters.

| Example | The putative reasons | A reason for *A* to *φ* | A reason for *A's φ*ing | A reason *A* had to *φ* | *A's* reason for *φ*ing |
|---|---|---|---|---|---|
| Award | My friend won an award | ✓ | ✓ | ✓ | ✓ |
| | I read that she had won an award in the newspaper | ✓ | × | ✓ | ✓ |
| | I believed that she had won an award | ✓ | × | ✓ | ✓ |
| Sally | A bear was chasing Sally | × | × | × | × |
| | She heard a bear-like sound | ✓ | × | ✓ | ✓ |
| | She believed that a bear was chasing her | ✓ | × | ✓ | ✓ |
| Eva | Her husband was unfaithful | × | × | × | × |
| | It appeared to her as though her husband was kissing another woman | × | × | × | × |
| | She believed that he had been unfaithful | ✓ | × | ✓ | ✓ |
| Climber | Loosening his grip would rid him of danger | ✓ | ✓ | ✓ | × |
| | It appeared as though loosening his grip would rid him of danger | ✓ | × | ✓ | × |
| | He believed that loosening his grip would rid him of danger | ✓ | × | ✓ | × |

**Table XVI-4:** Other examples for explanatory rationalism

## A.2 Explaining *in the right way*

Throughout this chapter, I assumed that if some fact is mysteriously related to the fact that the agent did what they did, then it explains it in the right way. This section provides more context to that assumption.

According to explanatory rationalism, something is the reason for which an agent does some action only if it is a reason for them to do it[9] and it explains why they did it *in the right way*. But what does it mean to say that it explains it *in the right way*?

Well, recall the case of the climber who was so unnerved by his resolution to drop his friend, that he loosened his grip on his friend unintentionally (i.e. not for a reason).[10] According to explanatory rationalism there was a reason for him to loosen his grip, which was, *inter alia*, the fact that he believed that doing so would rid him of danger.[11] Moreover, given that it causes him to loosen his grip, there is perhaps a (causal) sense in which the fact that he believed that loosening his grip would rid him of danger explains why he loosened his grip. That being so, a reason for him to loosen his grip explains why he loosened his grip, but it was nonetheless not his reason for loosening his grip.[12] Why not? Because it does not explain why he loosened his grip *in the right way*.[13]

What is the 'right way' of explaining? As I suggested in § (XIV), the *right way* of explaining an agent's action is not *merely* causal. To elaborate: while the right way might be *partly* causal (I will take no view on that), it is not *merely* causal. In particular, I have suggested that if a fact explains why an agent acted *in the right way* then it is mysteriously related to the fact that the agent acted as they did.[14]

What these remarks establish is that a mystery relation between a reason there is to act and the fact that the agent acted as they did is *necessary* for the reason to explain their action in the right way. For the purpose of this discussion, I assumed that it is *sufficient*; so that if a reason that an agent has to act is mysteriously related to the fact that they acted as they did, then it is their reason for acting. This assumption does not have a significant impact on my argument; it just makes the exposition less involved.

---

[9] I.e. it explains why it is *pro tanto* rational for them to do it.

[10] See § (XIV)4.

[11] It was a reason for him to loosen his grip because it explains why it was *pro tanto* rational for him to loosen his grip.

[12] This is clearly implied by the fact that he does not loosen his grip for a reason.

[13] Cf. 'What distinguishes actions which are intentional from those which are not? The answer that I shall suggest is that they are the actions to which a certain sense of the question 'Why?' is given application; the sense is of course that in which the answer, if positive, gives a reason for acting.' (Anscombe 1957, § 5)

[14] That is, I have suggested that the difference between a fact that *merely* causally explains why an agent acted and one that explains why they acted *in the right way* is that it is only in the latter case that the fact is mysteriously related to the fact that the agent did what they did.

## A.3 The problem of reasons for acting

Unlike any other theory of reasons (to my knowledge) explanatory rationalism distinguishes between the kind of reason picked out by the expressions 'a reason *to* act' and 'a reason *for* acting'. It insists that while a reason to act is just anything that explains why an agent's action is *pro tanto* rational, a reason for acting is something that is both a reason to act *and* counts in favour of the agent's doing it. A consequence of this view is that something could be a reason for an agent to do some action without being a reason *for* their doing it.

While this may sound paradoxical, in what follows I will argue that this is actually the least worst response to a problem that affects all theories of reasons.

### A.3.1 What reason there was for Sally to run

To the extent that we accept that it is *prima facie* reasonable to claim that Sally's reason for running is that she heard a bear-like sound (i.e. (F6)) and that an agent's reason for acting must be a reason there was for them to so act (i.e. (F9)), it follows that the fact that Sally heard a bear-like sound (in a wood that she knew to contain bears) was a reason for her to run. Indeed, it seems to me, if you hear something that sounds like a bear in a wood that you know to contain bears, then that is a *good* reason for you to run. Thus, I suggest:

(R5) The fact that Sally heard a bear-like sound is a reason for her to run.

### A.3.2 Reasons to act and reasons for acting

As I have already noted, it seems reasonable to say that if something is a reason *to* do some action then it is a reason *for* doing it, thus:

(R6) For any *p*, if *p* is a reason for *A* to *φ* then *p* is a reason for *A*'s *φ*ing.

### A.3.3 Reasons for acting count in favour of actions

In § (I)1.3, I noted that all reasons for acting count in favour of the actions for which they are reasons and that all reasons against some action count against doing that action. In § (I)4.4, I argued further that a reason for some action counts in favour of that action in virtue of its being a reason *for* it; that is, I argued that the 'for' preposition on its own is enough to give it that meaning.[15] Thus:

(R7) For any *p*, if *p* is a reason for *A*'s *φ*ing then *p* counts in favour of *A*'s *φ*ing.

[15] In § (I)4.4, I argued further that a reason for some action counts in favour of that action in virtue of its being a reason *for* it; that is, I argued that the 'for' preposition on its own is enough to give it that meaning – I left it open whether or not a reason for acting *also* counts in favour of an action in virtue of its being a reason. Similar remarks apply to reasons *against* some action.

### A.3.4 Counting in favour of

As I have already argued,[16] the most natural interpretation of what it is to count in favour of some action is to make it, in some respect, worth doing.[17] Thus:

(R8) For any *p*, *p* counts in favour of *A*'s *φ*ing if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing.

### A.3.5 What make's Sally's running worth doing

As I established in § (II)4, Sally's running is, in no respect, worth doing, so nothing makes it worth doing. And, in particular:

(R9) The fact that Sally heard a bear-like sound does not make her running, in any respect, worth doing.

### A.3.6 The Reasons for Acting Problem

The Reasons for Acting Problem is this: the fact that Sally heard a bear-like sound is a reason for her to run, so it is a reason for her running, so it counts in favour of her running, so it makes her running, in some respect, worth doing, but it doesn't make her running in any respect worth doing! Explicitly, the following claims are mutually inconsistent:

(R5) The fact that Sally heard a bear-like sound is a reason for her to run.

(R6) For any *p*, if *p* is a reason for *A* to *φ* then *p* is a reason for *A*'s *φ*ing.

(R7) For any *p*, if *p* is a reason for *A*'s *φ*ing then *p* counts in favour of *A*'s *φ*ing.

(R8) For any *p*, *p* counts in favour of *A*'s *φ*ing if and only if *p* makes *A*'s *φ*ing, in some respect, worth doing.

(R9) The fact that Sally heard a bear-like sound does not make her running, in any respect, worth doing.

This is a problem that *all* theories of reasons face (note how *all* of these claims are *prima facie* reasonable, although they are mutually inconsistent). The response of favourists and deliberativists to this problem is to reject (R5). Psychologism about reasons for acting rejects (R7).[18] Kearns and Star (2008) interpret what it is to 'count in favour of acting' as 'being evidence that one ought to so act,' so they would presumably reject (R8). I know of no one

[16] Recall the discussion of this point in § (I)4.2 in particular.
[17] Where what it is for an action to be worth doing is to be determined, except for the claim that whether or not some action is worth doing for some agent is independent of an their cognitive states.
[18] Psychologists would probably also reject (R5).

who would reject (R9). For reasons that I have set out at various points over the previous chapters, I don't find any of these options palatable.

In contrast, explanatory rationalism rejects (R6). I take this to be the least worst of the options available. Does it sound odd to say that something could be a reason to act without being a reason for acting? Somewhat. However, I think that the oddness of saying this dissipates when particular examples are considered: it does not sound odd to me to say that the fact that I believe that it is raining is a reason for me to take an umbrella but not a reason *for* taking an umbrella.

## A.4 Acting *pro tanto* rationally

According to explanatory rationalism, it is possible that an agent could do some action for a reason even though that action was merely *pro tanto* rational, and not all things considered rational. That is, explanatory rationalism suggests that even when an agent does one thing, despite judging something else to be all things considered worth doing instead, they still do what they do for a reason. Is this right?

Well, consider the following example:

> A man walking in a park stumbles on a branch in the path. Thinking the branch may endanger others, he picks it up and throws it in a hedge beside the path. On his way home it occurs to him that the branch may be projecting from the hedge and so still be a threat to unwary walkers. He gets off the tram he is on, returns to the park, and restores the branch to its original position… It is easy to imagine that the man who returned to the park to restore the branch to its original position in the path realizes that his action is not sensible. He has a motive for moving the stick, namely, that it may endanger a passer-by. But he also has a motive for not returning, which is the time and trouble it costs. In his own judgement, the latter consideration outweighs the former; yet he acts on the former. In short, he goes against his own best judgement. (Davidson 2004, 172 & 174)

The man in Davidson's example does act for a reason: he thinks that the stick may endanger a passer-by. It seems to me that the fact that the action is 'less than fully rational',[19] does not impinge upon its having been done for a reason.[20] But this means that we should be careful about saying that if one acts for a reason then they act rationally. Here is what I suggest:

- An agent acts *pro tanto* rationally if and only if their reason for doing it is a *pro tanto* reason to do it (i.e. something that explains why their action is *pro tanto* rational).

[19] To use Parfit's (2011, 34) term.

[20] Others are also of this view, for instance: 'The incontinent man holds one course to be better (for a reason) and yet does something else (also for a reason).' (Davidson 2001b, 34); '*Akrasia*, or weakness of the will, occurs when, in the face of conflicting reasons for and against *X*-ing someone makes an all-things-considered judgement that he ought not to *X*, but *X's* anyway and does so for a reason, namely, for whatever the reason in favour of *X*-ing was (which was included in the basis of the all-things-considered judgement).' (Hurley 1992, 130)

- An agent acts all things considered rationally if and only if their reason for doing it is an all things considered reason to do it (i.e. something that explains why their action is all things considered rational).

## A.5 Acting for a normative reason and knowledge

Several theorists hold that an agent acts for a normative reason only if they know it.[21] In § (XIV)2, I argued that a mystery relation between the fact that *p* and the fact that an agent believes that *p* is a necessary condition on the agent's knowing that *p*. Now, suppose further that it is also a *sufficient* condition (i.e. if the fact that *p* is mysteriously related to the fact that an agent believes that *p* then the agent knows that *p*).

If that were true, then explanatory rationalism would provide us with an account of *why* knowing a normative reason should be necessary for acting on it: because it is only if an agent knows a normative reason that it can mysteriously explain their action – and it is only their reason for acting if it mysteriously explains their action. Let me elaborate.

A normative reason to do some action must be mysteriously related to the agent's belief in it if it is to explain either why it was *pro tanto* rational for the agent to do that action, or (in the right way) why they did it[22]; that is, a normative reason to do some action must be mysteriously related to the agent's belief in it if it is to be the agent's reason for acting. And if we suppose that the mystery relation is sufficient for knowledge, then a normative reason will only be mysteriously related to an agent's belief in it if the agent knows that normative reason. So this is why an agent can only act for a normative reason if they know it: because if they don't know it then, *inter alia*, it won't explain why they did it *in the right way*, and it won't explain why it was *pro tanto* rational for them to do it.

[21] (E.g. Unger 1978; Hyman 1999, 2015; Hornsby 2008; McDowell 2013)
[22] Excluding weird cases (e.g. where a normative reason to do some action also happens to be a feature of the agent's psychology, which anyway explains why the agent's action is *pro tanto* rational).

# Bibliography


Alvarez, Maria. 2010. *Kinds of Reasons*. Oxford University Press.

———. 2013. 'Explaining Actions and Explaining Bodily Movements'. In *Reasons and Causes: Causalism and Anti-Causalism in the Philosophy of Action*, edited by Giuseppina D'Oro, 141–59. History of Analytic Philosophy. Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmiilan.

———. 2016a. 'Reasons for Action: Justification, Motivation, Explanation'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2016. http://plato.stanford.edu/archives/sum2016/entries/reasons-just-vs-expl/.

———. 2016b. 'Reasons for Action, Acting for Reasons, and Rationality'. *Synthese*, January, 1–18.

Anscombe, G. E. M. 1957. *Intention*. Harvard University Press.

Audi, Paul. 2012a. 'A Clarification and Defense of the Notion of Grounding'. In *Metaphysical Grounding*, edited by Fabrice Correia and Benjamin Schnieder, 101–21. Cambridge: Cambridge University Press.

———. 2012b. 'Grounding: Toward a Theory of the in-Virtue-of Relation'. *The Journal of Philosophy* 109 (12):685–711.

Audi, Robert. 2001. *The Architecture of Reason: The Structure and Substance of Rationality*. Oxford ; New York: Oxford University Press.

Broome, John. 2006. 'Reasons'. In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith, 29–55. Oxford University Press on Demand.

———. 2013. *Rationality through Reasoning*. The Blackwell/Brown Lectures in Philosophy 4. Chichester, West Sussex ; Malden, MA: Wiley Blackwell.

Chudnoff, Elijah. 2011. 'What Should a Theory of Knowledge Do?' *Dialectica* 65 (4):561–79.

Collins, Arthur W. 1997. 'The Psychological Reality of Reasons'. *Ratio* 10 (2):108–23.

Comesaña, Juan, and Matthew McGrath. 2014. 'Having False Reasons'. In *Epistemic Norms*, edited by Clayton Littlejohn and John Turri, 59–79. Oxford University Press.

Correia, Fabrice. 2010. 'Grounding and Truth-Functions'. *Logique et Analyse* 53 (211):251–279.

Crane, Tim. 2000. 'Dualism, Monism, Physicalism'. *Mind & Society; Heidelberg* 1 (2):73–85.

Dancy, Jonathan. 2000. *Practical Reality*. New York: Oxford University Press.

———. 2004. 'Two Ways of Explaining Actions'. In *Agency and Action*, edited by John Hyman, 25–42. Cambridge: Cambridge University Press.

———. 2006. 'Acting in the Light of the Appearances'. In *McDowell and His Critics*, edited by Cynthia Macdonald and Graham Macdonald, 121–34. Blackwell Publishing Ltd.

———. 2008a. 'On How to Act Disjunctively'. In *Disjunctivism*, edited by Adrian Haddock and Fiona Macpherson. Oxford University Press.

———. 2008b. 'Arguments from Illusion'. In *Disjunctivism*, edited by Alex Byrne and Heather Logue, 116–35. The MIT Press.

———. 2011. 'Acting in Ignorance'. *Frontiers of Philosophy in China* 6 (3):345–57.

———. 2012. ‘Response to Mark Schroeder’s “Slaves of the Passions”’. Edited by Mark Schroeder. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 157 (3):455–62.

———. 2014. ‘On Knowing One’s Reason’. In *Epistemic Norms*, edited by Clayton Littlejohn and John Turri, 80–96. Oxford University Press.

Darwall, Stephen. 2003. ‘Desires, Reasons, and Causes’. *Philosophy and Phenomenological Research* 67 (2):436–443.

Davidson, Donald. 2001a. ‘Actions, Reasons and Causes (1963)’. In *Essays on Actions and Events*, 2nd ed, 3–21. Oxford : New York: Clarendon Press ; Oxford University Press.

———. 2001b. ‘How Is Weakness of the Will Possible? (1969)’. In *Essays on Actions and Events*, 2nd ed, 21–42. Oxford : New York: Clarendon Press ; Oxford University Press.

———. 2001c. ‘Intending (1978)’. In *Essays on Actions and Events*, 2nd ed, 83–102. Oxford : New York: Clarendon Press ; Oxford University Press.

———. 2001d. ‘A Coherence Theory of Truth and Knowledge (1987)’. In *Subjective, Intersubjective, Objective*, 137–58. Oxford University Press.

———. 2004. ‘Paradoxes of Irrationality (1982)’. In *Problems of Rationality*, 169–88. Oxford : New York: Clarendon Press ; Oxford University Press.

Davis, Wayne A. 2003. ‘Psychologism and Humeanism’. *Philosophy and Phenomenological Research* 67 (2):452–59.

———. 2005. ‘Reasons and Psychological Causes’. *Philosophical Studies* 122 (1):51–101.

Everson, Stephen. 2009. ‘What Are Reasons for Action?’ In *New Essays on the Explanation of Action*, edited by Constantine Sandis, 22–47. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

FIDE. n.d. ‘Fide Laws of Chess Taking Effect from 1 July 2017’. World Chess Federation. https://www.fide.com/component/handbook/?id=207&view=article.

Fine, Kit. 2012. ‘Guide to Ground’. In *Metaphysical Grounding*, edited by Fabrice Correia and Benjamin Schnieder, 37–80. Cambridge: Cambridge University Press.

‘For, Prep. and Conj.’ n.d. *OED Online*. Oxford University Press. http://www.oed.com/view/Entry/72761.

Garrard, Eve, and David McNaughton. 1998. ‘Mapping Moral Motivation’. *Ethical Theory and Moral Practice* 1 (1):45–59.

Gettier, Edmund L. 1963. ‘Is Justified True Belief Knowledge?’ *Analysis* 23 (6):121.

Gibbons, John. 2010. ‘Things That Make Things Reasonable’. *Philosophy and Phenomenological Research* 81 (2):335–61.

Gjelsvik, Olav. 2007. ‘Are There Reasons to Be Rational?’ *Homage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, 142–147.

Goldman, Alvin I. 1967. ‘A Causal Theory of Knowing’. *The Journal of Philosophy* 64 (12):357.

———. 2012. *Reliabilism and Contemporary EpistemologyEssays*. Oxford University Press.

Hall, N. 2004. ‘Two Concepts of Causation’. *Causation and Counterfactuals*, 225–276.

Harman, Gilbert H. 1970. ‘Knowledge, Reasons, and Causes’. *The Journal of Philosophy* 67 (21):841–55.

Hecht, Stephen S. 2006. 'Cigarette Smoking: Cancer Risks, Carcinogens, and Mechanisms'. *Langenbeck's Archives of Surgery* 391 (6):603–13.

Heuer, Ulrike. 2004. 'Reasons for Actions and Desires'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 121 (1):43–63.

Hieronymi, Pamela. 2011. 'XIV—Reasons for Action'. *Proceedings of the Aristotelian Society (Hardback)* 111 (3pt3):407–427.

Hofweber, Thomas. 2009. 'Ambitious, Yet Modest, Metaphysics'. In *Metametaphysics: New Essays on the Foundations of Ontology*, edited by David John Chalmers, David Manley, and Ryan Wasserman, 347–83. Oxford ; New York: Clarendon Press.

Hornsby, Jennifer. 2007. 'Knowledge in Action'. *Action in Context*, 285–302.

———. 2008. 'A Disjunctive Conception of Acting for Reasons'. In *Disjunctivism*, edited by Adrian Haddock and Fiona Macpherson, 244–61. Oxford University Press.

Hurley, S. L. 1992. *Natural Reasons: Personality and Polity*. Oxford University Press.

Hyman, John. 1999. 'How Knowledge Works'. *The Philosophical Quarterly (1950-)* 49 (197):433–51.

———. 2011. 'Acting for Reasons: Reply to Dancy'. *Frontiers of Philosophy in China* 6 (3):358–68.

———. 2015. *Action, Knowledge, and Will*. Oxford University Press.

Kearns, Stephen, and Daniel Star. 2008. 'Reasons: Explanations or Evidence?' *Ethics* 119 (1):31–56.

———. 2009. 'Reasons as Evidence'. In *Oxford Studies in Metaethics*, edited by Landau. Vol. 4.

Kim, Jaegwon. 1988. 'Explanatory Realism, Causal Realism, and Explanatory Exclusion'. *Midwest Studies In Philosophy* 12 (1):225–39.

———. 1989. 'Mechanism, Purpose, and Explanatory Exclusion'. *Philosophical Perspectives* 3:77–108.

———. 1993. *Supervenience and Mind: Selected Philosophical Essays*. Cambridge Studies in Philosophy. New York, NY, USA: Cambridge University Press.

———. 2003. 'Blocking Causal Drainage and Other Maintenance Chores with Mental Causation'. *Philosophy and Phenomenological Research* 67 (1):151–76.

———. 2008. *Physicalism, or Something near Enough*. 3rd print., And 1st paperback print. Princeton Monographs in Philosophy. Princeton: Princeton Univ. Press.

Korcz, Keith Allen. 2015. 'The Epistemic Basing Relation'. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2015. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2015/entries/basing-epistemic/.

Kroedel, Thomas, and Moritz Schulz. 2016. 'Grounding Mental Causation'. *Synthese* 193 (6):1909–23.

Lehrer, Keith. 1971. 'How Reasons Give Us Knowledge, or the Case of the Gypsy Lawyer'. *The Journal of Philosophy* 68 (10):311–13.

Leuenberger, Stephan. 2014. 'Grounding and Necessity'. *Inquiry* 57 (2):151–74.

Lewis, David. 1987. 'Causal Explanation'. In *Philosophical Papers Volume II*, 314–240. Oxford University Press.

Littlejohn, Clayton. Forthcomming. ‘How and Why Knowledge Is First’. In *Knowledge First*, edited by A. Carter, E. Gordon, and B. Jarvis. Oxford University Press.

———. 2012. *Justification and the Truth-Connection*. Cambridge University Press.

Locke, Dustin. 2015. ‘Knowledge, Explanation, and Motivating Reasons’. *American Philosophical Quarterly* 52 (3).

Lord, Errol. 2008. ‘Dancy on Acting for the Right Reason’. *Journal of Ethics and Social Philosophy* 2 (3):1–7.

———. 2010. ‘Having Reasons and the Factoring Account’. *Philosophical Studies* 149 (3):283–96.

———. 2015. ‘Acting for the Right Reasons, Abilities, and Obligation’. In *Oxford Studies in Metaethics, Volume 10*, edited by Russ Shafer-Landau, 26–52. Oxford University Press.

Markovits, J. 2010. ‘Acting for the Right Reasons’. *Philosophical Review* 119 (2):201–42.

———. 2011. ‘Why Be an Internalist about Reasons?’ In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, 256–80. Oxford University Press.

McCain, Kevin. 2012. ‘The Interventionist Account of Causation and the Basing Relation’. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 159 (3):357–82.

McDonnell, Neil. 2015. ‘The Deviance in Deviant Causal Chains: The Deviance in Deviant Causal Chains’. *Thought: A Journal of Philosophy* 4 (3):162–70.

McDowell, John. 1982. ‘Criteria, Defeasibility, and Knowledge’. *Proceedings of the British Academy* 68:455–79.

———. 2013. ‘Acting in the Light of a Fact’. In *Thinking About Reasons*, edited by David Bakhurst, Brad Hooker, and Margaret Olivia Little, 13–28. Oxford University Press.

Mele, Alfred R. 2007. ‘Reasonology and False Beliefs’. *Philosophical Papers* 36 (1):91–118.

Mill, John Stuart. 1863. *Utilitarianism.* The Electric Book Company. http://www.myilibrary.com?id=124074.

Mitova, Veli. 2015. ‘Truthy Psychologism about Evidence’. *Philosophical Studies* 172 (4):1105–26.

———. 2016. ‘Clearing Space for Extreme Psychologism about Reasons’. *South African Journal of Philosophy* 35 (3):293–301.

O’Brien, Lilian. 2015. ‘Beyond Psychologism and Anti-Psychologism’. *Ethical Theory and Moral Practice* 18 (2):281–95.

Oliver, Alex. 1996. ‘The Metaphysics of Properties’. *Mind* 105 (417):1–80.

Olson, Jonas, and Frans Svensson. 2005. ‘Regimenting Reasons’. *Theoria* 71 (3):203–214.

O’Shaughnessy, Brian. 1973. ‘Trying (as the Mental" Pineal Gland")’. *The Journal of Philosophy*, 365–386.

———. 1980. *The Will: A Dual Aspect Theory*. Cambridge [Eng.] ; New York: Cambridge University Press.

Parfit, Derek. 2001. ‘Rationality and Reasons’. *Exploring Practical Philosophy: From Action to Values*, 17–39.

———. 2011. *On What Matters: Volume One*. Oxford University Press.

Pautz, Adam. 2010. 'Why Explain Visual Experience in Terms of Content?' In *Perceiving the World*, edited by Bence Nanay, 254–309. Oxford University Press.

Plantinga, Alvin. 1993. *Warrant: The Current Debate*. Oxford University Press.

Pollock, John L., and Joseph Cruz. 1999. *Contemporary Theories of Knowledge*. Rowman & Littlefield.

Raz, Joseph. 1999a. *Engaging Reason: On the Theory of Value and Action*. Oxford; New York: Oxford University Press.

———. 1999b. *Practical Reason and Norms*. Oxford University Press.

———. 2009. 'Reasons: Explanatory and Normative'. In *New Essays on the Explanation of Action*, edited by Constantine Sandis, 184–202. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

———. 2011. *From Normativity to Responsibility*. Oxford University Press.

Rosen, Gideon. 2010. 'Metaphysical Dependence: Grounding and Reduction'. In *Modality: Metaphysics, Logic, and Epistemology*, edited by Bob Hale and Aviv Hoffmann, 109–36. Oxford University Press.

Ruben, David-Hillel. 2004. *Explaining Explanation*. London: Routledge, Taylor & Francis e-Library.

———. 2008. 'Disjunctive Theories of Perception and Action'. In *Disjunctivism: Perception, Action, Knowledge*, edited by Adrian Haddock and Fiona Macpherson, 227–42. Oxford University Press.

Russell, Bertrand. 1917. *Mysticism and Logic, and Other Essays.* London : G. Allen & Unwin,.

Ryle, Gilbert. 1949. *The Concept of Mind*. London: Hutchinson.

Sandis, Constantine. 2012. *The Things We Do and Why We Do Them*. London: Palgrave Macmillan UK.

———. 2013. 'Can Action Explanations Ever Be Non-Factive?' In *Thinking About Reasons*, edited by David Bakhurst, Brad Hooker, and Margaret Olivia Little, 29–49. Oxford University Press.

Saporiti, Katia. 2007. 'A Seeming Solution to a Seeming Puzzle in Explaining Action'. *Action in Context*, 303.

Scanlon, Thomas. 1998. *What We Owe to Each Other*. Cambridge, Mass.: Belknap Press of Harvard University Press.

———. 2014. *Being Realistic about Reasons*. Oxford University Press.

Schaffer, Jonathan. 2009. 'On What Grounds What'. In *Metametaphysics: New Essays on the Foundations of Ontology*, edited by David John Chalmers, David Manley, and Ryan Wasserman, 347–83. Oxford ; New York: Clarendon Press.

Schnieder, Benjamin. 2011. 'A Logic for "Because"'. *The Review of Symbolic Logic* 4 (03):445–65.

Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford; New York: Oxford University Press.

———. 2008. 'Having Reasons'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 139 (1):57–71.

Setiya, Kieran. 2007. *Reasons without Rationalism*. Princeton, N.J.: Princeton University Press.

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

———. 1994. *The Moral Problem*. Philosophical Theory. Oxford, UK ; Cambridge, Mass., USA: Blackwell.

———. 1998. 'The Possibility of Philosophy of Action'. In *Human Action, Deliberation and Causation*, edited by Jan Bransen and Stefaan Cuypers, 17–41. Kluwer Academic Publishers.

———. 2004. 'The Structure of Orthonomy'. In *Agency and Action*, edited by John Hyman, 165–94. Cambridge: Cambridge University Press.

Sosa, Ernest. 2015. *Judgment and Agency*. Oxford University Press.

Stout, Rowland. 1996. *Things That Happen Because They Should: A Teleological Approach to Action*. Oxford Philosophical Monographs. New York: Oxford University Press.

———. 2009. 'Was Sally's Reason for Running from the Bear That She Thought It Was Chasing Her?' In *New Essays on the Explanation of Action*, edited by Constantine Sandis, 48–61. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

———. 2010. 'Deviant Causal Chains'. In *A Companion to the Philosophy of Action*, edited by Timothy O'Connor and Constantine Sandis, 159–65. Oxford, UK: Wiley-Blackwell.

Stoutland, Frederick. 1998. 'The Real Reasons'. In *Human Action, Deliberation and Causation*, edited by Jan Bransen and Stefaan E. Cuypers, 43–66. Philosophical Studies Series 77. Springer Netherlands.

———. 2007. 'Reasons for Action and Psychological States'. *Action in Context*, 75.

Turri, John. 2009. 'The Ontology of Epistemic Reasons'. *Noûs* 43 (3):490–512.

Unger, Peter. 1978. *Ignorance: A Case for Scepticism*. Oxford University Press.

Vogelstein, Eric. 2012. 'Subjective Reasons'. *Ethical Theory and Moral Practice* 15 (2):239–57.

Way, Jonathan, and Daniel Whiting. 2016. 'Reasons and Guidance (Or, Surprise Parties and Ice Cream)'. *Analytic Philosophy* 57 (3):214–35.

———. 2017. 'Perspectivism and the Argument from Guidance'. *Ethical Theory and Moral Practice* 20 (2):361–74.

Wedgwood, Ralph. 2002. 'Internalism Explained*'. *Philosophy and Phenomenological Research* 65 (2):349–69.

Whiting, Daniel. 2014. 'Keep Things in Perspective: Reasons, Rationality and the A Priori'. *Journal of Ethics and Social Philosophy* 8:i.

Williams, Bernard. 1981. 'Internal and External Reasons'. In *Moral Luck: Philosophical Papers 1973-1980*, 101–13. Cambridge University Press.

Yablo, Stephen. 2008. 'Wide Causation'. In *Thoughts: Papers on Mind, Meaning, and Modality*, 275–306. Oxford University Press.