Actual Control: Demodalising Free Will

David Christian Heering

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Philosophy, Religion and History of Science

September 2020

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

The right of David Christian Heering to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

The work in Chapter 3 of the thesis has appeared in publication as follows:

- Heering, David (forthcoming). Actual Sequences, Frankfurt-Cases, and Non-Accidentality. *Inquiry: An interdisciplinary Journal of Philosophy*.

The work in Chapter 2/4 of the thesis has appeared in publication as follows:

- Heering, David (2020). Intentionen, Misserfolg und die Ausübung von Fähigkeiten: Bemerkungen zu Agents' Abilities von Romy Jaster. *Zeitschrift für Philosophische Forschung, 74* (3), 454-459.

For my parents

For Rita Schumacher

And for the Losers of Botany House

Acknowledgments

I was fortunate enough to be able to draw on three sources of guidance for this thesis: Ulrike Heuer, Pekka Väyrynen (who took over for Ulrike in my second year), and Helen Steward. I owe all three of them the deepest gratitude possible. In the many garbled attempts to capture the thoughts expressed in the thesis, I gave them ample opportunity to be dismissive about my project. Yet, they never took it. Instead they have been nothing but encouraging and supportive, and their incisive comments have indispensably shaped and improved the thesis in a myriad of ways. I am especially grateful to Helen Steward, whose approach to philosophy has had a lasting influence on my own thinking, for her championship for my project. I am also very thankful to Maria Alvarez and Daniel Elstein for agreeing to be my external examiners and reading the thesis in such a short window of time.

A legion of people have contributed to the flourishing of both my thesis and me as a human being while writing it. First of all, I am thankful to my parents, who have always made the simplest and yet most essential contribution: their unconditional love and support. The same goes for my brother and his 'bounty family'.

I am grateful to the friends I have made in Leeds, chief among them the irreplaceable Adina Covaci and Pei-Lung Cheng, who have imbued with meaning and value what otherwise would have been (let's be honest) a rather pointless 4 years spent obsessing over a personal folly. Others deserving of mention include Alex Bréhier-Stamatiadis, Thomas Brouwer, Andreas Bruns, Arthur Carlyle, Aleksander Domoslawski, Will Gamester, Jessica Isserow, Tadhg Ó Laoghaire, Olof Leffler, Sam Mason, Alice Murphy, Emily Paul and Diana Sofronieva. My deepest thanks go to Miriam Bowen.

Thanks to my Berlin philosophy friends and colleagues, especially Niklaas Tepelmann, Daniele Bruno, Razvan Sofroni, Roland Krause, Berit Braun, Katharina Nagel and Stephanie Elsen. I also want to thank Romy Jaster and Barbara Vetter, as well as Thomas Schmidt and Leonhard Menges, in whose seminar on moral responsibility the ideas of this thesis first took (crooked) shape back in 2013.

A special nod goes towards Katharina Goebel and her Berlin WG for giving me refuge, Anastasija Bräuniger, who coerced me into writing theatre pieces during the PhD, and Linus Lutz for his kinship and for putting things in perspective (special thanks to Silvio Martin and our lockdown theatre reading group).

I owe gratitude also to Emily Herring, Hannah Robson, and Oliver Engley, and their incomparable green floor. If the ideas of this thesis had a colour, it would be green-floor green.

Thanks also to Alice Franzon, Ghada Habib, and Clara von Schwerin, who have independently been invaluable friends, but who have together provided me with an inexhaustible supply of awkward internet dating profiles they found. Your effort has kept me sane during the writing-up phase.

My work was made possible by The University of Leeds Research Scholarship (years 1-3), the Jacobsen Trust, and the Aristotelian Society Bursary (year 4).

Special thanks to Jana Doudova for her despite all unwavering support.

Thesis Abstract

Plausibly, agents act freely iff their actions are *responses* to reasons. But what sort of relationship between reason and action is required for the action to count as a *response*? The overwhelmingly dominant answer to this question is *modalist*. It holds that responses are actions that share a *modally robust* or secure relationship with the relevant reasons.

This thesis offers a new alternative answer. It argues that responses are actions that can be *explained by reasons* in the right way. This *explanationist* answer comes apart from the modalist answer. For it holds that actions are responses to reasons if they are explained by those reasons even if they don't share a modally robust relationship. Explanationism thus offers a novel way of vindicating the intuition that alternative possibilities don't matter to responding to reasons and (consequently) free agency.

The key dialectical position the thesis develops is that both modalism and explanationism constitute attempts to capture he core *type* of relationship encoded by the notion of a response. Responses to reasons, at core, involve a *non-accidental relationship* between reason and action. We can either understand non-accidentality as a modal phenomenon – as a modally robust tracking between two facts. Or we can understand non-accidentality as an explanatory phenomenon – as a special explanatory relationship between two facts.

According to my rival explanationist proposal, two facts share a non-accidental relationship iff we can give a unified explanation of why both obtain. Unified explanations are explanations of why [p&q] that cannot be decomposed into two (or more) separate independent explanations of p and of q. Consequently, according to explanationism about responding to reasons, actions are responses to reasons iff those reasons offer a rational explanation of the action that cannot be decomposed into separate independent components.

Contents

"I believe one should trust problems over solutions, intuition over arguments, and pluralistic discord over systematic harmony. Simplicity and elegance are never reasons to think that a philosophical theory is true: on the contrary, they are usually grounds for thinking it is false...If arguments or systematic theoretical considerations lead us to results that seem intuitively not to make sense, or if a neat solution to a problem does not remove the conviction that the problem is still there...then something is wrong with the argument and more work needs to be done...Superficiality is as hard to avoid in philosophy as it is anywhere else. It is too easy to reach solutions that fail to do justice to the difficulty of the problems. All one can do is try to maintain a desire for answers, a tolerance for long periods without any, an unwillingness to brush aside unexplained intuitions, and an adherence to reasonable standards of clear expression and cogent argument."

Thomas Nagel, Mortal Questions, x-xii.

General Introduction

It is part and parcel of our self-conception as rational agents that we have the capacity to respond to reasons. Sometimes, that is, there are good reasons for us to act in one way rather than another. We have the capacity to see those reasons and to base our actions on them. Moreover, this capacity is essential in the bold endeavour to continue to coexist and collaborate as human beings on a finite planet. For our relationship with actions that count as manifestations of the capacity to respond to reasons is markedly different from our relationship with other types of behaviour. An agent who bases their actions on reasons – rather than, say, impulse or whim – can make their actions explicable in a special way both to themselves and to others. They can give explanations – and sometimes defences – of their actions in terms of the reasons there were for so acting. And they thereby make their actions eligible for a variety of evaluative practices central to human interaction and communication.

It is also part and parcel of our self-conception as agents that we sometimes act freely. When we do, there is a special sense in which these actions are *ours*. They are actions that we can account for – and we are in turn held accountable for them. To some of our behaviour we are helpless bystanders. Free agents, by contrast, are the stewards of their actions. Free actions are actions under the agent's control.

This thesis will take for granted an idea about the connection between these two truisms about agency. The idea is that our actions are free *because* and to the extent to which they are responsive to reasons. Or in other words: free agency is grounded in an agent's *orthonomy* – their ability to think and act on the basis of what is overall the right thing to think and do.

Although no defence of this idea will be undertaken in this thesis, it is worth pointing out how naturally and seamlessly it combines the two aspects of human agency described above. There is a significant overlap between those actions for which no explication in terms of reasons is forthcoming and those actions we would treat as unfree. There is an overlap, in other words, between actions that we treat as ineligible for rational evaluation and actions over which we have lost ownership. The idea of grounding freedom in orthonomy is that this is because to lose ownership over an action *is* to lose the ability to account for it in terms of reasons – to lose the ability to give a rational explanation of why we did what we did. We are all, I assume, familiar with the phenomenon of 'snapping',

as we are apt to call it. States of heightened emotion – fits of rage and jealousy –can have a blinding influence on our sense of what is right. They can compel us into actions we significantly dissociate from. These interferences are most palpable when they concern the delicate relations we maintain to other agents. A perceived slight can ruin an evening. Even if we know no harm was meant by it, we may just not be able to let it go, shocked by our own persistence in the matter. These episodes are typically accompanied by a familiar sense of loss of control and ownership, in which it is as if all we can do is stand by and watch ourselves spiral. But they are also accompanied by an inability to account for our actions in a rational way. All we can do later is shake our heads regretfully and wonder what possessed us. Here, loss of control over our actions is intimately linked with the inability to account for them rationally. It seems like the lack of freedom and accountability for these actions just consists in not being able to explain them rationally – to ourselves and to others. They are untethered from the normative features of a situation that give shape and guidance to our behaviour. Or in other words: they fail to manifest the agent's capacity of orthonomy – more precisely, their responsiveness to reasons.

The fundamental question the thesis is occupied with is *what it is to respond to reasons*. The question is: When an action counts as a response to a reason – when it is based on that reason and therefore explicable in terms of it – what relationship does it have with that reason? This is, at base, the question about *the nature of the orthonomous relationship*. And it is one of the questions at the root of practical philosophy. It festers underneath debates about the nature of rationality, it informs accounts of the nature of action and agency, and it bubbles to the surface whenever philosophers deal with notions that involve the exercise of our orthonomy – such as 'knowledge', 'rule-following' or 'moral worth' –, notions that is, that at their core require the agent to respond to some feature of reality (truth, rules, moral rightness respectively for the examples).

The central feature of orthonomous relationships is that they are *robust* or *secure* in the sense of being *non-accidental*. When the agent responds to a reason, then it isn't just a fluke that they did what they had reasons to do. It is instead an agential achievement in which reason and action are non-accidentally connected.

Against this backdrop, the philosophical landscape has for the longest time known only one basic answer to the question about the orthonomous relationship – the modalist answer.

Modalism is the view that we can understand non-accidentality in terms of a kind of *modal security*. On the modalist picture, the reason and the action stand in an orthonomous relationship iff they are modal companions – they occur together or track

each other across a predefined set of alternative possibilities. This picture of orthonomous relationships is pervasive across many philosophical subdisciplines. It is at the core of safety conditions on knowledge in epistemology, conditions which dictate that beliefs count as knowledge only if they could not easily have been false. It is a silent presupposition in debates about the nature of moral worth. And it is the dominant view on how to spell out responsiveness to reasons. The view is so widespread, in fact, that it has led to an equally widespread neglect of the question about the nature of the orthonomous relationship. Why should we care about the question, after all, if an adequate answer is already available?

In this thesis, I develop an alternative answer to the question about the orthonomous relationship – the explanationist answer.

Explanationism is the view that we can understand non-accidentality as an explanatory phenomenon. On the explanationist picture, the action and the reason stand in an orthonomous relationship iff the reason explains the action in the right way. That is, according to the explanationist, we can gain an understanding of the orthonomous relationship by attending to the special rational explicability of our actions that reasons afford us.

The thesis argues that modalism is fundamentally mistaken and that explanationism is correct. The overall argument thus divides into two broad strategies. First, I argue that modalism is severely flawed. In fact, I think the thrall it has held over philosophy is responsible for some of the most pervasive problems orthonomy notions have faced in the last decades. Second, I offer a proper replacement for the modalist picture to counteract the continued conviction by many philosophers that there is no way to model the robustness of orthonomous relationships other than in terms of modal security. The worry is that actuality is too impoverished, as it were, to provide the kind of robustness needed. But the worry is wrongheaded. Actuality provides more than enough resources to understand the robustness/non-accidentality of orthonomous relationships – we just have to know where to look.

These two components of the thesis form the proper parts of what is the titular *demodalising* project – the project to erase all appeals to 'alternative possibilities' from accounts of orthonomy (and thereby free agency). In this sense, the project of this thesis is to formulate a new way of understanding the sentiment that alternative possibilities are irrelevant to free agency. What matters to free agency, according to this sentiment, is not what could have happened, but what did happen. According to the explanationist rendering of the sentiment, what matters to free agency is that we can account for our

actions in terms of reasons in the actual world, irrespective of how we would have acted under different circumstances.

The project will accordingly unfold in two steps. First, I will look at how modalist treatments of orthonomy continue to be beset by a deep structural problem (chapters 1 and 2) and how this problem is linked to/generalises to non-accidentality (chapter 3). I will then develop my alternative explanationist account in two layers. On the first layer (chapters 4 and 5), I focus on the idea that the achievement in basing one's actions on reasons is the manifestation of a competence. More precisely, I will hold that to respond to a reason is to *exercise* the capacity to respond to reasons. I will show how an exercise-based account of responding to reasons can account for a variety of ways in which agents respond to reasons. Most importantly, I will address questions about how cases of error fit into the orthonomy picture. If free agency is a matter of responding to reasons, how do we deal with cases in which agents seem to *fail* to respond to reasons and yet seem free? My answer is that in cases of error the agent can still account for their action in terms of their exercise of the capacity to respond to reasons. Thus, they still count as free. They have exercised their capacity to respond to reasons *badly*, I shall say. There is much controversy attached to this claim, and it will be addressed in due time.

On the second layer (chapters 6 and 7), I focus on the idea that exercising a capacity is itself an explanatory notion. To say that the agent exercised their capacity to respond to reasons is to say that they can account for their action in terms of the exercise of that capacity in a special way. They can give an *exercise-explanation*, as I shall say. The second layer spells out how these explanations work, and in doing so develops an explanationist account of the notion of non-accidentality. The core idea of this account is that we can understand non-accidentality in terms of a special form of explanation – *unified* explanation. Unified explanations are explanations that cannot be decomposed into separate independent component explanations. Those special rational explanations of our actions in terms of reasons, it will turn out, have exactly this structure. They are unified explanations. This picks up on a subplot in the thesis in general, namely that orthonomy concepts are non-decomposable (or 'prime') in a certain sense. The fact that S responded to the reason that p is not decomposable into the fact that p caused S to act and the fact that p made S's action rational, for example. Some crucial well-known problems with orthonomy notions are linked to this peculiar feature and can even be helpfully reframed using my explanationist approach.

The thesis is ambitious. In an attempt at intellectual honesty, and to prevent disappointment, let me briefly mention some of the many omissions and weaknesses fostered – but not excused – by this ambition. First, I will often use causal vocabulary somewhat permissively. Many things that taint a purist's causal ontology will be said to

cause or bring about happenings in my language. Second, some traditional issues that one might expect from a work on free agency are notoriously absent – such as any mention of the compatibilism/incompatibilism debate. I am lax about these issues because I believe that for the most part they don't matter to my project.

There will also be omissions of issues that *do* matter, such as issues over whether and in what sense alternative possibilities can be involved in explanation. I will address these omissions as avenues of future research in my general conclusion. Here, I will just say that whatever was omitted was omitted to make room for more programmatic aspects of my project. For above all, what this thesis hopes to do is to start the research program of thinking about orthonomy not in terms of alternative possibilities but in terms of explanations. The project is in a sense about *making things visible*. In areas of philosophy where the default analysis of a notion has become conventional wisdom, alternative pathways can disappear altogether. If this happens, the default will increasingly look like the *only* conceptual option – and all problems with it will consequently present as irresolvable. It is therefore paramount to make alternative options for analysis reappear by staking claims and laying a groundwork – irrespective of the exact details of what account will eventually rest on that edifice. It is this kind of project this thesis is ultimately engaged in.

Chapter Overview

Part 1: Problems with the Modalist Account of Responding to Reasons

Ch. 1 I introduce the idea that the orthonomy capacity crucial to free agency is reasons-responsiveness (RR). I discuss the view that reasons-responsiveness requires the ability to do otherwise traditionally conceived (as an ability to actualise open pathways into the future). I then introduce Frankfurt-cases as refuting this traditional view. I develop a preliminary understanding of responsiveness to reasons without the ability to do otherwise, which overcomes some initial anxieties about the concept. These steps are preparatory to setting up the discussion in chapter 2.

Ch. 2 I argue that state of the art views of reasons-responsiveness cannot overcome Frankfurt-like cases. State of the art views of reasons-responsiveness hold that Frankfurtian interveners are *masks* – they prevent an agent from exercising, but not possessing the right RR capacity. Consequently, there is a sense in which the agent remains reasons-responsive in Frankfurt-cases, a sense grounded by alternative possibilities in which interferences are absent. I argue that there are highly local intrinsic interferences that cannot be removed from the relevant alternative possibilities – *rational blind spots*. These cases point to a general structural problem for state of the art RR views that can only be solved if we develop an account of free agency in terms of the *exercise* of a *highly specific* RR capacity, an account that does away entirely with alternative possibilities.

Ch. 3 I argue that the problems encountered in chapter 2 generalise to a wide variety of notions. All of these notions have to do with achievement and orthonomy, and alternative possibilities seem irrelevant to all of them. I propose that this is because they all require non-accidental connections and the notion of non-accidentality can be understood in explanatory as opposed to modal terms. Thus, a new line for a demodalising project opens up. We can remove alternative possibilities from reasons-responsiveness by removing them from the underlying non-accidental relationship between reason and action.

Part II: The Explanationist Account of Responding to Reasons

Ch. 4 I focus on what it is to exercise a capacity. I argue that, in general, exercising the capacity to φ does not entail exercising the capacity to φ successfully. There may be exercises of the capacity to φ that are unsuccessful φ-ings. Hence, there may be exercises of the capacity to respond to reasons that are unsuccessful responses to reasons – false beliefs about what reasons there are, and actions performed for what were falsely believed to be reasons.

The argument is simple. Certain standards of success apply to a φ-ing only when that φ-ing counts as the exercise of the capacity to φ. If it was true that to exercise the capacity to φ is to φ successfully, then there could not be any φ-ings that fail those standards. For any *un*successful φ-ing would not count as an exercise of the capacity to φ at all.

Ch. 5 I argue for a view according to which responding to reasons is a unified notion across cases of error and cases of success. That is, there is only one relation involved in both cases of error and cases of success – that of exercising a capacity. I develop an account of responding to reasons from this insight. The account understands both steps in a full response – 'having' or 'possessing' of a reason and acting 'for' a reason – as characterizable through the notion of an exercise. I also argue that the competitor view – the view that responding successfully and responding unsuccessfully are two separate relations – does not cut across cases in the right way.

Ch. 6 I develop an explanationist account of non-accidentality. According to this account, coincidence and accidentality arise when we cannot answer a specific type of why-question – coincidence questions. A coincidence question asks about the co-instantiation of two (or more) facts p and q that share a salient similarity. In cases of coincidence, we can give an explanation of p and we can give an explanation of q, but we can't explain them *together*. Explanations in which we can instead explain the relational fact [p&q] are *unified*. In cases of non-accidentality, the explanation needs to be *maximally unified*.

I also give a general argument against modalism based on these considerations. The argument is this: Non-accidentality facts are *prime*. They cannot be decomposed. But modalism gives a composite account of them, according to which non-accidentality facts [p&q] decompose into a range of mere co-instantiations across modal subspaces.

Ch. 7 I conjoin the results of chapters 5 and 6 by developing an account of what it is to exercise a capacity in terms of how the corresponding exercise-explanations work. Exercise-explanations pick out unified

explanatory structures, which is why the notion of an exercise automatically picks out non-accidental connections.

I conclude with an overview of what types of metaphysics might underpin the explanatory structures the thesis focuses on. The gist is that Humeanism – the view that there are no necessary connections between distinct existences – is ill-suited to support non-accidental connections, because it offers a composite picture of the universe according to which all there ultimately is are separate independent particulars.

Chapter 1:

Orthonomy, Reasons-Responsiveness, and the Ability to Do Otherwise

1. Introduction

This thesis offers a new way of understanding the claim that free agency is grounded in an agent's orthonomy. It offers an understanding of the claim which is compatible with the tenet that free agency can only be grounded in features of the actual world, and never in alternative possibilities of any sort.

But in order to understand the dialectical context in which this offer is made, I need to do some set-up. Accordingly, the primary purpose of this chapter is to introduce (some of) the crucial ideas that this thesis will grapple with in their most general form, as well as the distinctions and positions that form the dialectical backdrop for my arguments in later chapters. In particular, the chapter is concerned with the role that the 'ability to do otherwise' plays in the idea that free agency is a matter of orthonomy – the ability to get things right. The ability to do otherwise is widely regarded as crucial to free agency, but it does not integrate well with other ideas often closely linked to the orthonomy idea – most importantly the Frankfurtian (as in Harry Frankfurt) claim that what matters to free agency is what happens in the actual world, not what would have happened under different circumstances. The chapter explores, and develops a working definition of, an idea of orthonomy now commonly appealed to in the literature that does not rely on 'the ability to do otherwise' traditionally conceived. However, in this new conception of orthonomy, alternative possibilities are not entirely eliminated, they merely play a different role. I will argue in the next chapter that this constitutes a major problem for the idea of orthonomy, which required radical correction. The rest of the thesis then provides this correction.

Before we get ahead of ourselves however, let us explore the idea of orthonomy and its relation to alternative possibilities now.

2. Orthonomy

I said that this thesis investigates the idea that free agency is grounded in an agent's capacity to respond to reasons. This is an ability to adjust their thinking and acting to normative significance in the world[1] – their *orthonomy* (Wolf 1990, Pettit and Smith 1990, Pettit and Smith 1993, Pettit and Smith 1996, Smith 2003).

Orthonomy, most broadly understood, is the idea that as rational agents, we have the capacity to think and act in line with how we ought to think and act – the capacity to be orthonomous. As Pettit and Smith put it:

> To be orthonomous, as distinct from autonomous, an agent's evaluations and desires have to be sensitive to his recognition of normative requirements: reasons that may be offered in support of evaluative claims. To the extent that there are normative requirements to be satisfied, the achievement of orthonomy will therefore represent something distinct from any sort of internal harmonization; it will represent a way of coming into line with something outside the realm of desire: with the reasons in favor of the relevant evaluative claims […] Coming into line with the norms will require either a sensitivity to rationally binding reasons or attunement with the world. (Pettit and Smith 1996, 443)

Orthonomy is therefore a capacity to be *sensitive* to, or *responsive* to normativity – or perhaps we can say a capacity to *track* normativity. What exactly this responsiveness amounts to is mysterious, and finding a fitting account will be what large parts of this thesis are concerned with. But we can rely on our intuitions about responsiveness here. When agents possess and/or exercise these capacities, they will stand in orthonomous relationships with facts relevant to their conduct, relationships which we then describe with a range of orthonomy notions, such as knowledge, perception, action, moral worth, and responsiveness to reasons. I will explain why the latter notion works best for grounding free agency below.

Before I do, we need to get two preliminary clarifications out of the way.

First, it will be helpful to point out that there are two ways in which the basic idea that free agency is grounded in orthonomy can be understood. Orthonomy Views are focussed on capacities of agents, and a central feature of capacities is that agents may or may not exercise them. Accordingly, *Exercise Orthonomy Views* hold that an agents'

[1] The thesis is written in a realist spirit that I no longer share. I am now more convinced that an interesting constitutivist account of normativity can be developed from the account I will present in this thesis. The idea is that normativity in 'generated' by the exercise of agents' rational abilities (a Kantian and/or Korsgaardian idea). However, this narrative is hidden within the text of the thesis, which holds officially that reasons are mind-independent facts.

control is grounded in the agent's *exercise* of their relevant capacity. *Possession Orthonomy Views*[2] hold that an agents' control is grounded in their *possession* of the relevant capacity. According to the latter, but not the former, it will be enough for an agent to be free with respect to some φ-ing if, at the time of φ-ing they possessed the ability to φ orthonomously. According to the former, even if the agent possessed at the time of φ-ing the ability to φ orthonomously, it may still be false that they were in control of their φ-ing. What matters is whether in φ-ing they also exercised that capacity. The distinction will come up in chapter 2, but I flag it up here for anticipatory purposes. Due to matters of space, I will not be able to engage with an interesting hybrid category, developed in Wolf (1990) and Nelkin (2011). According to these views, the freedom-relevant capacity is *the ability to do the right thing for the right reasons*. When this ability is exercised, what matters is the exercise of the ability. When the agent fails to exercise it, i.e. does not do the right thing, what matters is the possession of the ability.[3] These hybrid views are therefore committed to treating the grounds of free agency asymmetrically. Freedom is grounded in the possession of orthonomy for 'bad' actions and it is grounded in the exercise of orthonomy in 'good' actions. I will argue in chapter 4 and 5 that we should treat agential failures as bad exercises of capacities. Thus, I will implicitly argue against the asymmetry thesis adopted by Wolf and Nelkin. Further, I will argue for an Exercise View (and against a Possession View) in the next chapter. In this chapter, I will assume a Possession Orthonomy View in the background for reasons of presentational ease.

Second, a clarification about the relationship between free agency and orthonomy. The idea that freedom is grounded in orthonomy is to be understood as a *metaphysical* proposal about those features *in virtue of which* agents possess the characteristic sense of ownership of their actions and attitudes associated with free agency (see Sartorio 2016a, ch.1).[4] It is the sense in which agents are creditworthy in any number of respects (moral, epistemic and so on) for their actions. Since this sense of an agent's actions and attitudes being *theirs* in a credit-relevant way focusses on the agent's relation to their actions, not merely their intentions, another way to express the orthonomy idea is via the notion of *control* (Fischer and Ravizza 1998). Free actions are under their agent's control in a credit-relevant way. The question is in virtue of *which* features agents control their actions in this way. The line of thinking I shall pursue here is the relevant grounding

[2] For explicit endorsements of the view, see Brink and Nelkin (2013), 292; Wallace (1994), 183.

[3] Some issues relevant to these asymmetrical views come up implicitly in chapter 4.

[4] I will not concern myself here with the metaphysical intricacies of the grounding relation. See Raven (2015) for an overview. All I require is that it is an ontological dependence relation which we latch onto with our 'in virtue of' and sometimes ur explanatory talk (as in 'agents are free *because* they are reasons-responsive').

feature is an agents orthonomy understood as their responsiveness to normative reasons.

Let me explain. A powerful motivation for thinking agents are in control in virtue of being responsive to normative reality is that almost all paradigms of unfree action/attitude involve a failure on the agent's part to be responsive to some feature of their normative situation. Delusional and compulsive agents are unfree – and our evaluative and legal practices link their lack of free agency to their incapacities to recognize and translate into action what matters normatively in their situation. We consider the paradigms of free agents, on the other hand, as agents who believe and act in possession of full cognitive capacity and in full recognition of the facts that make their actions and beliefs normatively recommended.

Importantly, the idea that free agency is grounded in attunement to normative reality is not just the idea that the agent's volitional system is in harmony with their values or deliberative systems. For even agents who count as autonomous in that their intentions and actions line up with what they want and value may be unfree. They may be severely deluded or out of touch with what matters. We care about autonomy only insofar as it expresses orthonomy, and we consider it unfit to ground free agency when it diverges from the standards of orthonomy (Watson 1975, Wolf 1990, Smith 2004). That is, the distinguishing feature of orthonomy accounts is that they not only require an internal coherence of the agent, but an external relationship with normative reality for the agent to be free.

What is normative reality? I will here follow the popular idea that we can express most or all aspects of what is meant by this term in the vocabulary of *objective normative reasons*. Normative reasons are considerations that speak in favour of or against some attitude or action. They are fact-like entities.[5]

The term is also used to designate *relations*.[6] This is because reasons are always reasons *for* or *against* actions or attitudes, and they are always reasons *for* agents to perform these actions or have these attitudes. However, I will mainly talk about reasons as those entities that occupy the reason-position in these relations, because I take it these entities are what agents are in tune with when they are in tune with normative reality.[7]

If the term 'normative reality' is understood to mean the landscape of considerations for and against actions and attitudes that agents face, then it makes sense to understand

[5] The idea was popularised by Raz (1975) and Scanlon (1998).

[6] See Raz (1975) and (1999); Dancy (2004a); Cuneo (2007); Skorupski (2010); Scanlon (2014)

[7] The issues that are raised by this distinction are implicitly discussed in chapter 5.

the orthonomy-relevant capacity as the capacity to respond to reasons – or *reasons-responsiveness* (RR[8]) (Wolf 1990; Wallace 1994; Fischer and Ravizza 1998, Nelkin 2011). The capacity is to be understood as composed of at least two important subcapacities: The epistemic capacity to recognize reasons and the practical capacity to translate reasons into action – that is, act for those reasons (Wallace 1994; Fischer and Ravizza 1998, Nelkin and Brink 2013).

These preliminary clarifications then give a more precise version of the idea that free agency is grounded in orthonomy: An agent's actions are free in virtue of the agent's possession of the capacity to recognize and act for reasons.

I will eventually argue (ch. 2 and 3) that this idea – as it is traditionally spelled out – clashes with another important tenet about free agency: that it is exclusively grounded in features of the actual sequence. In order to set up this argument, I now need to address how Orthonomy Views are related to the so called 'ability to do otherwise'.

3. Orthonomy and the Ability to Do Otherwise

A good way to start thinking about free agency – or control, as I shall say sometimes – is to think about the difference between compulsive and non-compulsive actions. I offer no theory of compulsion here. On the contrary, I think we can identify the relevant phenomenon by example.[9] Compulsive actions are those actions we feel internally coerced to perform. Their paradigm manifestations include addictions, and actions based on strong desires we significantly dissociate from (think compulsive hand-washing). While the actual empirical difference between such actions and ordinary actions might be overemphasized in philosophical idealisation, a crucial intuition that most of us can agree on is that even if there are no actual compulsive actions, in the possible world in which there are, these actions are paradigmatically unfree. For now, this intuition is all I need.

Imagine then a philosophically idealized compulsive agent, a severe addict who acts from the desire to take heroin. Call this agent *Compelled Cody* for reference purposes.

Contrast this agent with an agent who acts from a desire to try heroin because she is curious, call her *Curious Carmen*.

[8] I will use RR both as a noun and an adjective 'is reasons-responsive'. It will be clear from context in what sense it is used.

[9] In fact, large strands of philosophical literature rely on this pretheoretic understanding of the notion, none more than Watson's famous challenge to distinguish cases of compulsion from cases of mere weakness (see Watson 1977,324).

We think there is a difference between Curious Carmen and Compelled Cody. We think Carmen is and Cody isn't in control of taking the heroin. In accordance with the metaphysical project of grounding control, we can then ask: In virtue of what is Curious Carmen in control and Compelled Cody isn't?

It is tempting to give the following answer: While Curious Carmen has the ability to do otherwise than take the drug, Compelled Cody does not have this ability. Carmen, but not Cody, has metaphysical access to genuinely open branching pathways into the future.[10] It is up to her which of these she takes.[11] In contrast, it is not up to Cody which path he takes. He can't but follow his desire for the drug. Carmen's action is free then, because there is a host of agentially accessible alternative possibilities in which she does something other than take the drug. And Cody's action is unfree because these alternative possibilities do not exist for him. By alternative possibilities I just mean non-actual situations, that is, parts of possible worlds, and they are agentially accessible in the sense that agents have the power to render them actual by choosing to engage in a given course of action. Views that emphasise these alternative possibilities come with a specific analysis of control, which implies a strong modal dependence claim:

Control$_1$: An agent S is in control of their φ-ing because S has the ability to do otherwise than φ.

Entails

Strong Modal Dependence Claim: Control metaphysically depends on there being a set of agentially accessible alternative possibilities.

Control$_1$ is not incompatible with thinking that the relevant ability to do otherwise is the ability to respond to reasons. But it insists that part of the grounds of control, even when understood as a rational ability, is what the agent *would have done* in different circumstances. It insists that it is part of an agent's orthonomy that they could have done otherwise. It is a way of understanding an agent's orthonomy.

[10] Alvarez (2013) and Steward (2012) understand agency to be so-called 'two-way power' – the power to do or not do. This power is not equivalent with the ability to do otherwise traditionally conceived, because the second part of the two-way power – the 'not do' part – is not obviously tantamount to the existence of an open pathway into the future in the sense needed for the ability to do otherwise. It is, after all, a manifestation of the power of agency that consists in a non-doing (not an active omission, that is, but simply the absence of a doing). Because of intricacies related to how we understand this claim, I will not here discuss how two-way power views relate to Orthonomy Views.

[11] Whether or not this makes the world in which Carmen exists one in which determinism doesn't hold is not the topic of this thesis.

However, *control*$_1$ has traditionally sat uncomfortably with many Orthonomy Views. This is because it famously clashes with another class of cases in the vicinity, the kind of case that Harry Frankfurt (1969) introduced and that I will label FSC for Frankfurt-style case. FSC's go like this:

> **Frank**
>
> Frank is about to cast his vote in the 2016 presidential election. Unbeknownst to him, a manic Trump-supporter has installed a device in Frank's brain that monitors Frank's decision procedure. Should Frank show signs indicating an impending decision to vote for Clinton, the device will intervene and make Frank vote for Trump. But Frank has already made up his mind. He finds Trumps policies and general demeanour attractive and considers this sufficient reason to vote for Trump. So he votes for Trump on his own without showing signs indicating a contrary decision. Thus, the device remains inactive.

In virtue of the device, it seems true that Frank lacks the ability to do otherwise (than vote for Trump). He has no genuinely open pathways into the future that are up to him to choose. Yet most of us intuitively judge that Frank is in control. So the ability to do otherwise as I presented it so far is not what grounds control. Or so FSCs suggest. In other words, although there are no agentially accessible alternative possibilities for Frank, Frank is in control. So the strong modal dependence claim is false. So control$_1$ is false.

Views that follow Frankfurt's argument are commonly called *Actual-Sequence Views*[12], while views that reject Frankfurt's conclusion are called *Leeway Views*[13] (see Timpe 2012; Sartorio 2016a for the label). Actual-Sequence Views emphasize the importance of the actual history of an action to whether the agent has control over it. More precisely, they emphasize the parts of the actual history of the action that are explanatorily salient. Leeway Views emphasize alternative possibilities as crucial to control. However, a bit more care needs to be taken in comprehending the dialectical situation.

This is because technically FSCs involve two claims: one positive and one negative (Sartorio 2016a, 18). The positive claim holds that free agency is grounded in part in features of the actual sequence. This claim alone is not in contradiction with *control*$_1$ and the *strong modal dependence claim* because it is entirely possible that free agency is grounded *both* in actual-sequence features and in alternative possibilities. In fact,

[12] Major advocates of Actual-Sequence Views include Fischer (1982), (1994), (2006), and (2012); Fischer and Ravizza (1998); Frankfurt (1969) and (1971); Haji (1998); McKenna (2003), (2008), and (2013); Mele (1995), (2006); Hunt (2000), (2005); Sartorio (2016a), Pereboom (1995), (2001), (2014); Zagzebski (2000).

[13] van Inwagen (1983); Ginet (1990); Kane (1996) and many others.

several authors can be interpreted as holding such a hybrid view (see for example Kane 1985, p.59; van Inwagen 1983). Hybrid views of course still entail the *strong modal dependence claim*. The more controversial claim supported by FSCs is therefore the negative claim that free agency is *not* grounded in anything other than features of the actual sequence. It is this negative claim that is incompatible with the strong modal dependence entailed by *control*$_1$. The negative claim can be understood as a metaphysical constraint on the type of feature than can ground control. It holds that the only features fit to be grounds for control are actual-sequence features. That is, only the explanatorily salient features of an action's actual history are fit to be the grounds of an agent's control over it.

Together, the positive and the negative grounding claim can be expressed in the following slogan:

> **Frankfurt's Insight**: Control over an action is grounded exclusively in features of the actual sequence.

When I speak of *Actual-Sequence Orthonomy Views* in the following, I understand these approaches to be committed to *Frankfurt's Insight* (even if they fail to accommodate it). *Frankfurt's Insight* illustrates that FSCs can be represented as offering a contrastive thought. Two actions φ-ing and ψ-ing will differ with respect to whether they are free, the thought goes, only if they differ with respect to some property of the actual sequence leading towards the action. If two actions do not differ with respect to any such actual-sequence property, but differ with respect to a counterfactual property, then there will be no difference with respect to their freedom.[14]

FSCs have produced what is perhaps the most lucrative cottage industry known in philosophy, rivalling even that surrounding Gettier cases. Ironically, it is this labyrinthine cottage industry which makes it hard to assess their dialectical relevance – and which often inspires scepticism about their probative value.[15] The sheer amount of papers produced and rabbit holes to consider also make it an exhausting task to try to defend the cogency of FSCs. I will therefore not try.[16] Instead, I will take a more holistic stance

[14] This of course presupposes that two action can differ with respect to a counterfactual property without differing with respect to an actual-sequence property.

[15] I once heard a professor in an introductory lecture say that FSC are 'a dead end' because so much labyrinthine literature exists about them. In no other part of philosophy have I seen people so eager to take the fact that a debate is highly fine-grained as evidence that it has no philosophical worth.

[16] This is not to say that there aren't interesting and worthwhile parts of that literature. Here is a selection:

1. Helen Steward (2009) has pointed out, there may be different rationales for adopting a principle that asserts the importance of the ability to do otherwise (a PAP) in the first place. In my vocabulary, there are normative motivations usually to do with intuitions about fairness, and there are metaphysical motivations, usually to do with intuitions about agency (chapter 3 argues that the core intuitions are about 'non-accidentality'). One interesting strand of fairness-based counter to FSCS is to resettle with a PAP that describes an ability directly

towards them. What I will try to establish later (ch. 3)[17] is that FSCs should not be considered a phenomenon endemic to the free will debate. They are the manifestation of a much deeper clash between intuitions concerning a key conceptual requirement common to all orthonomy notions (such as knowledge, moral worth, perception, and others). As such, they are merely one species of a much wider ranging genus of cases that comes up in a lot of philosophical subdisciplines. In a nutshell, they are part of a class of cases which feature what I shall call *modal exploits* – cases that show that a relationship may still be non-accidental even though there is no modal tracking between the relata (see chapter 3, section 4; chapter 6, section 7 and 8). Modal exploit cases are widespread and include cases of finking and masking in the literature on dispositions as well as cases of luck in epistemology. If FSCs are therefore incoherent or lack probative value, this cannot be because of some detail germane to the free will debate. And the fact that this wider genus of case is so prevalent is a prima facie reason to think that it points to *some* issue common to the relevant subdisciplines. Philosophers aren't merely confused or inattentive about some detail in counterfactual reasoning, when they hold that FSCs are telling. Instead, what FSCs point to is that orthonomy notions can be understood in two radically different ways, and we should see Frankfurt-*type* cases as a manifestation of the irreconcilability of these two ways. In the end, we can still disagree

normatively relevant, for example the ability to avoid blame (Otsuka 1998; Moya 2007 and Wyma 1997). Note that it is plausibly true in FSCs that the agent does have the ability to avoid blame. In my opinion, these attempts should ultimately be understood as the project of uncoupling blameworthiness from free agency.

2. The most familiar counterstrategy to FSCs is the 'dilemma defense' (see Widerker 1995; Ginet 1996; and Kane 1996). The argument goes like this: What is the relationship between the agent's alternative action/choice and the (involuntary) prior sign the intervener will use to spring into action? If it is a deterministic relationship, then FSCs are claiming an agent can be determined yet free, which is going to be rejected by incompatibilists. If the relationship is indeterminstic, then we are saying that the agent might have decided otherwise without the sign occurring – so their actual choice was not unavoidable. Fischer (2010) has forcefully argued against the first horn of this dilemma. An interesting array of improved FSCs have been developed against the second (Stump 1996 and 2003, Haji 1998, Pereboom 2000, 2001and 2014, Hunt 2005). The debate is unhelpfully muddled due to not clearly adhering to Steward's distinction explained in 1. I think the second horn of the dilemma can be countered by pointing out that it does not follow from the fact that the agent might have done otherwise that they have the ability to do otherwise. They still lack this ability even in cases of indeterministic connections between prior sign and choice.

3. Alvarez (2009a) and Steward (2008), (2009) argue that the ability to do otherwise (or at the very least the ability to not do whatever the agent actually does) is indispensable for our understanding of agency as opposed to free agency. So FSCs clash with a more fundamental concept, on this view.

4. Vihvelin (2013) classifies FSCs according to the method of intervention in 'bodyguard' versions and preemptor versions. In bodyguard versions, what would have triggered the intervention is any actional token, an attempt or the beginning of an action of the agent. In these cases, Vihvelin argues, the agent keeps the ability to *begin to choose* is therefore free with respect to the exercise of that ability. In preemptor versions, the intervention is triggered by an earlier non-actional token, like the agent blushing. In these, Vihvelin argues, the agent keeps the ability to choose otherwise, they have the global (Vihvelin: narrow) ability to choose otherwise (they choose otherwise in a sufficient proportion of intervener-free possible worlds) and nothing stands in their way of exercising this ability. A merely counterfactual intervener for Vihvelin is never an actual obstacle to exercising an ability the agent possesses. Sartorio (2016b) forcefully objects to both strategies. The issues also come up in an exchange between Fischer and Vihvelin in Vihvelin (2000), Fischer (2008), Vihvelin (2008). I have heard (but not yet seen in publication) from both Maria Alvarez and Helen Steward that a Vihvelin-type argument can be helpfully combined with the view that agency the is power to act or not act.

[17] I advance the same argument in a more condensed form in Heering (forthcoming).

about these ways to understand orthonomy (chapter 6, section 8 offers an argument as to why Frankfurt-cases must succeed however). But this does not mean that Frankfurt-cases are incoherent or dialectically worthless. In fact, if FSCs are indeed a symptom of a deeper philosophical conflict about orthonomy, then their probative value is much higher than previously thought.

So I will here take it for granted that FSCs do establish Frankfurt's Insight and thereby undermine *control*$_1$.

If we cannot understand orthonomy as involving the ability to do otherwise, the question about the difference between Compelled Cody and Curious Carmen is still open. The foregoing has merely added a further constraint to it. The task is now to find a property that distinguishes both Curious Carmen and Frank from Compelled Cody. That is, the challenge is to spell out orthonomy without recourse to an agent's open pathways into the future.

More precisely, since Actual-Sequence Orthonomy Views are still committed to our starting thought that orthonomy is a kind of ability, they must give us an account of what *type* of ability orthonomy is if it *isn't* in part the ability to actualise one of several branching pathways into the future. In order to understand the answer typically given by Actual-Sequence Orthonomy Views, we need to understand some basic resources from the literature on abilities. The next section introduces these resources and points out some important choice points they open up for Actual-Sequence Orthonomy Views. Section 5 then applies these resources to reasons-responsiveness, dispersing some initial doubts about the notion in the process. What we should have at the end of the next two sections is an intelligible working definition of reasons-responsiveness sans ability to do otherwise as it is understood by Actual-Sequence Orthonomy Views. I will then carry this definition into chapter 2, where I will discuss state of the art improvements of it.

4. Orthonomy Without the Ability to Do Otherwise

4.1. Systems

Reasons-responsiveness is considered by Actual-Sequence Orthonomy Views as an ability – but not as the ability to actualise branching pathways into the future. To understand the alternative target orthonomy ability, a few more general remarks about capacities and abilities are in order. These remarks will not be exhaustive by a long shot, and they will in part rely on Romy Jaster's recent treatment of abilities in her *Agents'*

Abilities. I recommend this book for a full understanding of the issues in the background, which I lack the space here to address.

Abilities, as they are understood by Actual-Sequence Views, fall within the larger class of dispositional properties (which includes powers, dispositions and other potentialities)[18]. These properties are best characterised by what happens in sets of non-actual circumstances, in terms of what would have happened, that is, if certain conditions had obtained.[19] For example: A glass is fragile. This does not mean that it *does* break, but it means that it *could* break under certain conditions. What conditions? This question points to a first level of characterisation for dispositional properties. They are, in part, intrinsic structures of objects which react to certain triggering conditions. The non-actual circumstances relevant to whether object possesses a dispositional property are those circumstances in which those intrinsic properties are triggered. Consider again the fragile glass. Presumably, its fragility isn't just a brute fact. The glass is fragile because it has specific intrinsic properties (molecular structure is a plausible candidate). It breaks when hit with a strong enough force because these properties will react[20] to the force in a way that brings about what we can identify as the paradigmatic manifestation behaviour for fragility.

All this is philosophically standard for dispositional properties. I would like to add at this stage a less standard condition. It seems to me that the intrinsic bases for dispositions, capacities, and abilities should be considered as *organised structures*. It is not enough for an object to be fragile that it simply possesses properties ABC in just any arrangement. For in order for the glass to produce the typical manifestation behaviour upon encountering a large enough force, the properties that are in part causally responsible for the outcome have to interact in a number of specific ways. This is easier to see in a different example. Take my tachometer – the small device that measures the speed of my bicycle. This device has (or is) the disposition to measure speed. It has this disposition in virtue of a number of interrelated intrinsic properties that I would be able to list if I knew how a tachometer works. Importantly though, we cannot just rearrange these properties willy-nilly. The tachometer is a small machine, and machines work because their intrinsic parts are arranged in exactly the way that they are. This is what I mean when I say that the intrinsic properties that form the basis for disposition need to

[18] This is not to say that they are dispositions. See Vetter & Jaster (2017), Vetter (forthcoming).

[19] Franklin (2015) takes this to automatically entail that Actual-Sequence Orthonomy Views also hold that the ability to do otherwise is necessary for free agency.

[20] As indicated in the general introduction to this thesis, I am here permissive in my ascription of what types of things can stand in causal relations. This is because the causal level will for the most part only play a minor role in this thesis. The account I will develop is compatible with various views on causation (but not with all of them, see ch. 7., part II).

be organised structures. Like the tachometer's properties, they need to be functionally related, i.e. related in a way that enables the relevant base to reliably produce the relevant outcome. I will call such arrangements *systems*.[21]

Systems are the intrinsic bases of dispositional properties. It must be pointed out though that the external circumstances in which we place these systems also have effects on the truth conditions of dispositional ascriptions. Go back to the fragile glass. Now wrap the glass in a protective layer. Is it still fragile? Here is a description of the situation that I would hope transcends metaphysical disagreements. There is *a sense in which* the glass is still fragile, but there is also a sense in which it isn't. In the sense in which the protected glass keeps its fragility, we focus mainly on the underlying intrinsic system of the glass, mostly ignoring its current circumstances. In the sense in which the glass loses its fragility, we are taking its current circumstances into account. The same goes for abilities: A swimmer whose legs have been encased in concrete (they clearly knew too much!) in some sense lacks the ability to swim. But she has not lost the intrinsic system that grounds that ability. So, in another sense, she keeps the ability to swim. I shall refer to these senses as 'local' and 'global', following Whittle (2010) – local abilities being those that agents lose in inauspicious circumstances and global abilities being those they keep.[22] The next chapter will address whether we should understand the claim that free agency is grounded in the capacity to respond to reasons as a claim about a local or a global capacity. Here, I just raise the general distinction.

The foregoing remarks are part of the search for a proper working definition of reasons-responsiveness as interpreted by Actual-Sequence Orthonomy Views. In order to arrive at such a definition, some more stage setting is required. In particular, I need to address the modal analysis of capacities usually favoured by Actual-Sequence Orthonomy Views.

4.2. Counterfactualism and Possibilism

Two types of analysis of dispositional properties can be considered widespread in philosophy, both of which make use of other types of modal vocabulary to illuminate ability ascriptions.[23]

[21] There may be exceptions to this, such as fundamental dispositions such as mass and charge (if these indeed are dispositional properties). But those exceptions don't matter for my purposes here.

[22] The idea is labelled and understood slightly differently by different authors. Berofsky (2002) distinguishes between 'type and token' abilities, Vihvelin (2013) distinguishes between 'narrow and wide abilities'. Maier (2014) and Jaster (2020) distinguish between 'general and specific' abilities. Austin (1956), 218 famously identifies 'all in' abilities.

[23] Although other accounts exist, such as in terms of habituals (Fara 2005) and in terms of options (Maier 2013).

Counterfactualism holds that true ability ascriptions can be captured in terms of ranges of counterfactuals about the agent. For example, an archer has the ability to hit the bullseye, according to these approaches, iff she *would have hit in bullseye had she been in the right circumstances.* 'The right circumstances' is a placeholder for now. I discuss it below.

Unless Counterfactualism is developed in tandem with a different (for example a categorical) semantics for counterfactuals, it will be beholden to the commitments of the traditional Lewis-Stalnaker semantics of counterfactuals. According to the L-S semantics, a counterfactual like 'If S tried to hit the bullseye, S would hit the bullseye' is true iff S hits the bullseye in all close possible worlds in which S tries (i.e. the worlds in which the antecedent holds).[24] Consequently, to say that S has the ability to hit the bullseye in circumstances C will be equivalent to saying that S hits the bullseye in *all* close possible worlds in which C is the case.

Possibilism on the other hand holds that ability ascriptions are statements of restricted possibility. To say that someone can do something in the ability-sense, according to this idea, is to say that it is possible that they do that thing in view of a certain range of facts. Here is Lewis giving expression to the idea:

> To say that something can happen means that its happening is compossible with certain facts. Which facts? That is determined (...) by context. An ape can't speak a human language – say, Finnish – but I can. Facts about the anatomy and operation of the ape's larynx and nervous system are not compossible with his speaking Finnish. The corresponding facts about my larynx and nervous system are compossible with my speaking Finnish. But don't take me along to Helsinki as your interpreter: I can't speak Finnish. My speaking Finnish is compossible with the facts considered so far, but not with further facts about my lack of training. What I can do, relative to one set of facts, I cannot do, relative to another, more inclusive, set. (Lewis 1976, 149)

Since possibilism understands ability ascriptions as restricted possibility claims, it is committed to an existential statement. An archer has the ability to hit the bullseye, then, if there is at least one possible world of the relevant description in which the archer does hit the bullseye.

Both views have their own drawbacks and advantages. I am here mentioning them because an understanding of their difference is important to the hybrid analysis I will

[24] See Stalnaker (1968), Lewis (1973a, 1979).

favour in developing a working definition of reasons-responsiveness in the next section. Before I can get to that, one further stagesetting issue needs to be addressed.

4.3. Volitionalism and Non-Volitionalism

One distinction that is especially important for RR theories concerns the condition under which abilities are exercised. Dispositional properties manifest in response to certain stimuli. Glasses break upon collision with hard objects. Matches light when struck. But what are the conditions for the manifestation of abilities like reasons-responsiveness? One suggestion is that they are triggered by an appropriate volitional state of the agent, like an intention, or, as Vihvelin suggest, the agent's *trying* (Vihvelin 2013, 187). The proposal is given support by the view that agential abilities in general are triggered by volitional states. Jaster's agential abilities for instance are triggered by intentions. She does not however seem to count reasons-responsiveness as an agential ability (Jaster 2020, 284). Let me refer to the idea that reasons-responsiveness is triggered by volitional states as *Volitionalism*.[25]

Discussing triggering circumstances for reasons-responsiveness is important because these circumstances will determine the antecedents of the relevant counterfactuals (for counterfactualism) or the relevant class of worlds (for possibilism) relevant to the RR ability. If we follow the Vihvelin proposal, for example, a first rough draft for an RR ability will read something like: *S has the ability to respond to reasons iff, if S tried to respond to reasons, S would respond to reasons*. Or in the possibilist rendering: *S has the ability to respond to reasons iff there is at least one close possible world in which S intends to respond to reasons and succeeds*.

I will primarily work with a Non-Volitionalist version of reasons-responsiveness in the following. This is because I find that Volitionalism sits awkwardly with the idea that orthonomy is a kind of attunement to the normative landscape. Agents who are in tune with, capable of tracking and understanding, that is, some aspect of their environment are not usually able to flick this attunement on or off on command. Rather, they in a sense can't help but notice whatever they are attuned to. Take the following case: I am very good at picking up on social cues that indicate past or present romantic entanglements between people, however clandestine they might be. Though I would sometimes like this ability to be triggered by my volitional states, so as to have it remain dormant in socially inconvenient situations, I can't help but notice relationships – whether it fits my volitional states or not. This is because these sorts of attunement abilities are akin to

[25] Some might reject the idea that responsiveness to reasons is a *triggered* dispositional property outright. But I am here discussing it because major contribution in the debate have not shared this rejection.

perceptual capacities. And just like I can't help but see things, I can't help but notice romantic entanglements. The triggers of these abilities are the things themselves – or appearances resembling them – not an agent's volitional states. We should think of an agent's ability to respond to reasons in the same way – as the ability to pick up on normative significance in situations in which it is salient (or situations appearing as if it was). That is to say: the trigger for the ability to respond to normative reasons are normative reasons (or their indicators), not intentions or tryings. This is Non-Volitionalism. I will mark this aspect of my preferred version of RR by calling it a *capacity* rather than ability because the latter term carries Volitionalist connotations.

I should mention one important background issue for the choice point of Volitionalism vs. Non-Volitionalism. The issue is that the sense in which agents *can't help* but take into consideration normative significance might be taken to suggest that according to Non-Volitionalist Orthonomy Views, reasons are an external force, which effectively coerces agents into responsive action. It is essential to understanding Orthonomy Views that this is a nonsensical suggestion from their perspective. According to Orthonomy Views, free agency *consists* in a proper attunement to normative reality. Reasons are in this context not to be regarded as external coercing influences, but as a part of the osmotic equilibrium between agent and normative landscape in virtue of which agents are free. There is another important way to put this point. According to Orthonomy Views, there is no freedom-conferring quality in the ability to act *against* reasons – the ability to recognize and understand reasons, but to choose to act contrary to their demands. To Orthonomy Views, such actions are dysfunctional instances of agency, characteristic of the normative anchorlessness and irrationality that undermine rather than grant freedom (see Wolf 1990, especially pp. 55-62).

According to the Non-Volitionalist version of reasons-responsiveness (cast in a possibilist light), *an agent has the capacity to respond to reasons iff there is at least one relevant possible world in which there is a reason for the agent to* φ *and they recognize and act for that reason*.

This is a first approximation of a plausible candidate capacity description for the kind of responsiveness to reasons Actual-Sequence Orthonomy Views hold grounds free agency. But it needs elucidation, which I provide presently.

5. Simple Reasons-Responsiveness as Orthonomy

In the last section, we saw the basic structure for what the capacity to respond to reasons looks like if we don't want to understand it in terms of the ability to do otherwise. In this section, I want to fill in more details.

To stat this process, go back to our three example agents. What do Frank and Carmen have in common that Cody lacks?

At the core of the idea that free agency is grounded in orthonomy is the insight that Curious Carmen exhibits a special sort of rational flexibility with respect to her actions that Compelled Cody does not exhibit. Carmen, recall, is taking the heroin because she is curious about its effects, not because the desire to take it is overwhelming. If we keep this in mind, we can imagine a whole host of situations in which Carmen would refrain from taking the drug, for example if she had more urgent business or if taking the drug would severely harm her children, or if...etc. As several authors have noted, we should not understand the counterfactual claim in RR as made true by a single possibility. For that to be possible, it would have to be highly specified, which it is evidently not. Instead, we should understand the counterfactual as "a whole raft of possibilities", that is, as elliptical for a whole host of more specific counterfactuals. An agent is RR with respect to her φ-ing then, if she responds correctly to the present reasons in a whole host of scenarios. If there were sufficient reasons for φ-ing, she would φ. If there were sufficient reasons for not φ-ing, she would not φ. If there were sufficient reasons for φ*-ing, where this is an action very close in description to φ-ing, she would φ* and so forth. Smith (2003) demonstrates this concept with an agent who has the ability to respond to reasons for and against giving a correct answer to a question. The idea of "rafts of counterfactuals" is that the agent is RR if she would answer correctly in ever so slight variations of her circumstances, the question, the way it is asked etc. What we are doing is imagining situations in which sufficient reasons against taking the drug are present and Carmen does indeed act for those reasons. That is, we imagine changing the balance of reasons and we pay attention to whether this change will be reflected in Carmen's behaviour.

If we imagine Cody in those same situations, the fact that his desire is overwhelming is best accounted for by saying that he would take the drug anyway, even if sufficient reasons against taking the drug were present. The property that distinguishes Carmen from Cody is then her intact reasons-responsiveness.

Prima facie, Frank from FSCs fits into the reasons-responsiveness picture very well. Frankfurt-agents are, by stipulation, not psychologically or rationally impeded in any way. They decide on their own – and this locution is plausibly taken to mean that they decide based on their reasons for acting as they do. So it seems like Frank is similar to Carmen and dissimilar to Cody in that he is reasons-responsive. The elegance of this solution is why Actual-Sequence Orthonomy Accounts in terms of RR have seemed especially appealing to many.

In many ways, this idea is unclear and open to a host of worries that may severely diminish its attractiveness however. So let me briefly develop a more detailed account. I will do this by assembling the pieces for a plausible account of RR from the most obvious objections to the idea that control is grounded in RR.

The first worry might be that RR comes cheaply. After all, agents are responsive to a myriad of reasons for a lot of actions all the time. As I sit here now, I am aware that my enjoyment of movies is a reason to go to the Leeds Film Festival, while my upcoming deadline for supervision is a reason not to go. The same goes for a lot of other actions that seem abstract options for me right now. In a similar vein, even highly compulsive agents like Cody seem to be responsive to reasons pertaining to how they take their drug. It is true of them for example, that they would take the drug using a clean needle rather than a dirty one and that they would take a more potent version of the drug rather than a less potent one – if given those choices.

This kind of responsiveness seems to be a feature of human action in general and is therefore too broad to capture what is distinctive about free actions.

This worry points to two clarifications that we must apply. First, note that the notion of RR important to grounding control must be able to make true locutions like "S was in control of her φ-ing because she was RR". So the class of reasons relevant in this context is always the class of reasons relevant to φ-ing alone, not the class of reasons relevant to φ-ing and ψ-ing. For example, if we want to assess whether an agent is in control of her buying milk (for she may be an obsessive, unfree milk-buyer), the relevant reasons she needs to be responsive to are reasons for buying milk, not for her marrying Robert over Fred. The agent might be responsive to a lot of reasons for abstract options. Relevant to her control, on the RR picture, are only those reasons that favour/disfavour her actual decision/action. Second, importantly, this class of reasons is never merely the class of reasons *for* φ-ing. The normative landscape also includes considerations that *disfavour* courses of action. The relevant notion of RR must therefore include reasons *for and against* φ-ing. These considerations provide us with the first fragment for RR (relevant to control), which I cast here in a Counterfactualist light:

An agent S is RR with respect to her φ-ing iff S would successfully respond to reasons for and against φ-ing if they were present.

Of course this fragment still only offers a very broad notion of RR. Most notably, it is still true of this version of RR that compulsive agents are RR because they are, in a sense, responsive to reasons for and against their action. Compulsive Cody would take a higher dosage and use a cleaner needle, if those options were offered to him, which is certainly a way to respond to reasons.

In order to avoid this result, we must be attentive to the *kind* of reason that is important to RR. Compulsive Cody and Curious Carmen are different in their responsiveness, I said, in that Carmen, but not Cody, would respond to *sufficient*[26] *reasons* for and against taking the drug. This specification is certainly not true of Cody. Even if he does notice that taking the drug is a health hazard, he would take the drug anyway. By comparison, Carmen would refrain from taking the drug if she was (reliably) told it posed a serious health hazard. Note that this specification also gets rid of the cases where Cody responds to reasons pertaining to clean needles and stronger dosage. For those reasons are not sufficient reasons *to take the drug* by a long stretch. They are reasons to take the drug in a specific way, given that there is sufficient reason to take it in the first place. It makes little sense to say that an agent who has no reason to take the drug has nevertheless sufficient reason to take it in a specific way. If at all, an agent who has no sufficient reason to take the drug, has conditional reason to [if he has reason to take the drug, take it in a safe manner]. Responsiveness to this kind of reason is not part of the relevant notion of RR.

The foregoing considerations render the following result:

An agent S is RR with respect to her φ*-ing iff S would successfully respond to sufficient reasons for and against* φ*-ing if they were present.*

I take it that this dispenses with the general worry that responding to reasons comes cheaply and so cannot be used to pick out what is distinctive about free actions. There remain similar worries, however.

The current discussion may have brought to the fore already the worry that arises from the gulf between Possibilist and Counterfactualist rendering of the RR capacity. For notice that if we follow, as I have done, the standard semantics for counterfactual conditionals, the above version of RR holds of agents only if they respond successfully in *all* close worlds in which the balance of reasons is slightly different. Upon reflection, this will seem like an excessively perfectionist standard. No agent is a perfect responder. Even highly attuned agents will under some circumstances fail to see reasons relevant to their current action or fail to muster the willpower to translate them into action. Take Carmen, whom I have so far described as a fairly rational woman, in tune with reasons for and against taking the drug. But Carmen too might be susceptible to the appeal of the drug when described by a rhetorically gifted drug dealer even if there are good reasons no to take the drug. Hence, not all worlds in which the balance of reasons is

[26] A larger problem for all orthonomy views is raised in Ruth Chang's (2001a, 2001b, 2013, 2014, 2020) work. According to her 'Hypervoluntarism', normative reality sometimes (in fact routinely) underdetermines our choices. We therefore need a normative power to 'create reasons', as it were, to put our weight behind one of several options equally well supported by normative considerations.

different are worlds in which Carmen responds correctly. The relevant point is that this imperfection is not of the freedom-undermining sort. Carmen is still *pretty* responsive. She responds correctly in most of the paradigm situations. And we would not withdraw our judgment that she is creditworthy for and in control of her taking the drug if we learned that she does not respond to reasons in all relevant situations. Hence, the sense of RR in which RR plausibly grounds control is not yet captured by the current Counterfactualist reading of RR (at least assuming standard semantics).

It might seem like the Possibilist reading can help us out. Recall that according to Possibilism, we interpret the ascription of a capacity as a statement about restricted possibility. That Carmen possesses the capacity to respond to reasons then means that there is at least one possible situation in which the balance of reasons is different and Carmen acts differently. This does capture a sense of RR in which Carmen remains responsive despite not being a perfect responder.

However, as you can perhaps already see, the Possibilist reading is too weak. For presumably even Compelled Cody would refrain from taking the drug if there were overwhelming reasons not to take it. Imagine for example that Cody's drug den is on fire and Cody understands he will burn if he takes the drug now. That he will burn otherwise is an overwhelming reason not to take the drug and Cody does react to that reason if it is present. So even though Cody's responsiveness to sufficient reasons might be a bit more sluggish than that of Carmen, it is still true that he would react to sufficient reasons for and against taking the drug in virtue of the scenario just described.

Some middle ground has to be found here, evidently (Nelkin and Brink 2013 refer to this as 'the Goldilocks standard').

My solution to this problem follows the general solution to the parallel problem for abilities more generally (Jaster 2020). Abilities surely sometimes misfire. I have the ability to score a complicated kind of goal in table-soccer. But sometimes I fail. That is, in some of the worlds where I try, I fail. This does not rob me of my ability to score the goal though. However, if I score the goal in only a fraction of the worlds where I try, it is not true that I can score the goal in the sense of having the ability to do so. This suggests going for a middle ground: It is enough if I score the goal in a *suitable proportion* of scenarios where I try. We should think about RR along the same lines. It is too strict to require that the agent successfully responds to her reasons in all relevant worlds, but too lax to require that she is successful in only one possible world. So the plausible middle ground is to require that she is successful in a suitable proportion of worlds.

Now, it might be a thorny question what determines which proportion is suitable, but it seems to me the success rate of a severely addicted agent like Cody is clearly below any

sensible threshold. So, with the introduction of the proportion clause, we have found a way to respond to the worry about Cody's responsiveness in extraordinary scenarios. It might be true that Cody would not take the drug if sufficient reasons like the prospect of burning or dying or torture were present. Still, those cases are only a small fraction of the cases we use to ground the truth of RR. In most of those cases, in most ordinary cases in fact, sufficient reasons for refraining to take the drug are present, but Cody takes the drug anyway. Correspondingly, Carmen reacts according to the balance of reasons in most cases, given the way she was described. So we get the correct result for her too. This gives us:

> **RR**: An agent S is RR with respect to her φ-ing iff S successfully responds to sufficient reasons for and against φ-ing in a suitable proportion of the relevant possible worlds in which they are present.

RR contains the terminology of *relevant* possible worlds. This terminology is necessary because the trouble with failing to respond to reasons under certain conditions is far from over, as the next chapter will discuss. For the purposes of this chapter however, I hope that RR gives a good enough idea of the kind of dispositional property that Actual-Sequence Orthonomy Views typically use to advance their account.

The next chapter argues that RR is incapable of grounding free agency, and neither are the various versions of RR that have emerged from the recent literature. At base the worry will be that RR still relies too heavily on alternative possibilities. So while it does not obviously clash with *Frankfurt's Insight*, it does not adhere to its spirit. This leads to a deep structural problem for Actual-Sequence Orthonomy Views, which is most easily expressed in polemics: they aren't *actual* Actual-Sequence Views.

Chapter 2:

Complex Reasons-Responsiveness and the Recalcitrance of Actuality

1. The Problem for RR and Weak Modal Dependence

The previous chapter introduced the idea that free agency – or control over actions – consists in an agent's orthonomy – the capacity to get things right, and to get them right about normative matters. This chapter argues that the way this capacity and its relationship to orthonomy are traditionally conceived is mistaken. The traditional view is that the kind of reasons-responsiveness that grounds freedom consists in the possession of a special dispositional property – a property understood in terms of far-away alternative possibilities. My argument will be that this conception still gets into trouble with Frankfurt's original insight, the insight that what matters to freedom is the actual sequence. What grounds freedom instead is the exercise of a highly local capacity to respond to reasons – a capacity the agent has in the circumstances they are in and as the kind of agents they actually are.

To set up the views I will attack in this chapter, recall that my straightforward version of understanding this capacity to respond to reasons read like this:

> **RR**: An agent S is RR with respect to her φ-ing iff S successfully responds to sufficient reasons for and against φ-ing in a suitable proportion of the relevant possible worlds.

I also pointed out that Frankfurt-cases, if coherent, provide very strong reason for thinking that control does not depend on the *accessibility* of alternative possibilities to the agent. Frankfurt-cases falsify the strong modal dependence claim about control and modal properties. But RR does not entail such a strong modal dependence claim, so RR might be thought to be compatible with the insight of Frankfurt-cases.

However, while RR approaches are often committed to this consequence of FSCs, it is easy to see that they are committed to a weaker modal dependence claim. After all, RR clearly still counts as what we might call an essentially modal property. It is a property that can only be spelled out with reference to certain alternative possibilities. If this

property is metaphysically linked to control, there is an important sense in which control depends on alternative possibilities. That is, it seems that:

Control$_2$: An agent S is in control of her φ-ing because S is RR with respect to her φ-ing.

Entails

Weak Modal Dependence: Control over S's φ-ing metaphysically depends on there being a close[27] set of alternative possibilities.

RR is a less metaphysically demanding property than the original ability to do otherwise because it does not involve reference to the agent's metaphysical access to the alternative possibilities. It merely claims that *if* there were sufficient counter-reasons, the agent *would* react otherwise. No special power of bringing these alternatives about is allocated to the agent.[28]

However, here is the central problem for RR. It is, ironically, falsified by FSCs as well. To see this, take note of how RR above is spelled out. In order for Frank – the agent in Frankfurt-cases – to be reasons-responsive, it must be true that he would react differently in a suitable proportion of worlds where sufficient counter-reasons are present. But if it is true in virtue of the intervening device that he could not have done otherwise, then surely it is true that he could not have reacted otherwise as well. In particular, think about worlds where Frank takes into account Donald Trump's misogynist behaviour and tries to react to that reason. In those worlds, the implanted device will make Frank vote for Trump if he tries to respond to the misogyny reason. So because in worlds where sufficient counter-reasons are present, the device would or does intervene, it is false of Frank that he would have reacted differently had sufficient counter-reasons been present (Fischer and Ravizza 1998, Smith 2003, Sartorio 2016, as well as Mckenna 2013 identify this as a central problem as well).[29]

[27] I will explain the significance of this below.

[28] This idea is similar in many respects to the ideas of classical compatibilism (Hume, *Enquiry Concerning Human Understanding*, p.73; Ayer 1954; Hobart 1934, Smart 1961 but nowadays mostly associated with Moore 1912). The classical compatibilist idea is that we can analyse the ability to do otherwise in terms of conditionals like: 'If the agent wanted to, they could do otherwise.' The problem (Chisholm 1964, in Watson (ed.) 1982, pp.26–7; or van Inwagen 1983, pp. 114–9; most concisely summarised in Lehrer 1968) is that this generates a sense of ability which seems irrelevant to freedom – a global sense in my terms. This discussion is an important precursor to my argument in this chapter. For my argument will be that modern analyses of RR retain the core problem of classical compatibilist analyses of abilities.

[29] I discuss my non-accidentality-based solution in Heering (forthcoming).

This is the crucial problem for RR accounts. They want to claim that Frank is acting freely because he is reasons-responsive. But a straightforward account of RR renders the contrary result. It follows in addition that just as FSCs falsify the *strong modal dependence* claim, they also falsify *the weak modal dependence claim* because they feature agents who are in control but lack access to the relevant close alternative possibilities.

In response to this problem, state of the art reasons-responsiveness views will usually deploy resources familiar from the literature on dispositions to advance a more complex notion of reasons-responsiveness, a notion true in virtue of intervener-free alternative possibilities. This chapter argues that even these versions of RR clash with actual-sequence intuitions expressed in cases very much like Frankfurt cases. They ultimately clash with what I called Frankfurt's Insight in chapter 1, that is.

My aim is to show that this problem is not about finding the correct type of alternative possibility we use to spell our reasons-responsiveness. It is about understanding RR through alternative possibilities *at all*. This is a deep structural problem for orthonomy accounts. Orthonomy, it seems, is not about what the agent could have done, but what they *actually do*.

Understood according to the distinctions introduced in chapter 1, my conclusion for this chapter will be that only the agent's *exercise* of her very *local* capacity to respond to reasons for and against her actual action can ground control.

However, to understand the structural problem at the core of my argument, we first need to understand the resources used to construct complex reasons-responsiveness. I will introduce those resources now.

2. Abstracting Away

2.1. Interferences and Dispositions

In order to understand complex reasons-responsiveness, it will be helpful to first briefly look at some issues in the literature on abilities and dispositions.

A porcelain cup is protected by a powerful sorcerer. If it were to hit a hard object, the sorcerer would intervene and change its intrinsic structure to make it temporarily bouncy. Intuitively, the cup stays breakable even though it is not true that it would break if it hit a hard object. Such unusual interference is sometimes called a *fink* (Martin 1994; Lewis 1997). A glass wrapped in bubble wrap retains its fragility – its disposition to break if struck with sufficient force. But due to the bubble wrap, it seems false to say that the glass would break if it was struck with sufficient force. Such familiar interference is

sometimes referred to as a *mask* (Johnston 1992; Fara 2005). Finally, imagine a sturdy rock linked to a bomb such that if something were to strike the rock, the bomb would go off. In this case, although it is true that the rock would burst into pieces if it were hit by a hard object, we would be hesitant to say that the rock possesses the disposition of fragility. These interferences are referred to as *mimics*[30] in the literature.[31]

In all of these cases, an object retains a dispositional property even though some unusual actual circumstances make the counterfactual we would prima facie connect to that property false. This is partially due to the standard (Lewis-Stalnaker) semantics of counterfactual conditionals (henceforth just counterfactuals).[32] Note that in assessing the truth of counterfactuals the antecedents of which are not specified in any particular way, we are imagining possible worlds that are as close as possible to the actual world. In other words, the truth-conditions of counterfactuals are given by the set of closest possible worlds that satisfy the specifications of the antecedent of those counterfactuals. Hence, in each case I have just given, a simple counterfactual analysis of the type *o has disposition d to* φ *in response to stimulus S iff if o underwent S, o would* φ returns the wrong results. For the unspecified counterfactual will be true in virtue of the worlds closest to the actual world – worlds in which the relevant interferences become active.

However, the more particular perpetrator behind interference cases is the fact that dispositions require the circumstances of their manifestation to be *auspicious* (I adopt this parlance from Fisher 2013). The circumstances have to be conducive to the exercise of the relevant dispositions. Unless these circumstances are suitably specified in the antecedents of the relevant conditionals, the modal base of disposition ascriptions will be infected with too many cases in which circumstances are not conducive for manifestation.

Several versions of this idea exist in the literature, including the idea that auspicious circumstances can be understood in terms of conditions ideal for manifestation (Mumford 1998, 88–90), in terms of conditions typical of manifestation (Malzkorn 2000, 456–459), or typical for the trigger circumstances to occur (Fara 2005).[33] Perhaps most famously, David Lewis has suggested that, roughly, auspicious circumstances are those in which the bearer of the dispositions retains its intrinsic properties that, together with the trigger, will be a complete cause of the manifestation (Lewis 1997). There is also the idea that what we need is to provide a maximal specification of the conditions under

[30] Johnston (1992), Martin (1994), Lewis (1997), Bird (1998), Fara (2001) all present versions of mimicking cases.

[31] There are also so-called 'antidotes' (Bird 1998).

[32] See Stalnaker (1968), Lewis (1973a, 1979).

[33] There are also ''normal circumstances'' (Bird 1998, pp. 233–4), ''standard conditions'' (Gundersen 2002, p. 407), ''background conditions'' (Cross 2005, p. 324), and ''ordinary conditions'' (Choi 2009, p. 576).

which dispositions trigger in order to solve all interference counterexamples (see Lewis 1997, 144; Manley and Wasserman 2007). The idea is that if we specify precisely the force and angle of impact, and all other physical conditions which have to obtain for an object to break in the antecedent of our conditional, we will be able to exclude all intervention-worlds from our basis, because these are presumably worlds in which some of those conditions are violated. There is however now considerable consensus in the literature that such specification is impossible (Mumford and Wasserman 2008).

Going forward I will instead adopt a sort of abstraction of those approaches that reject the specification strategy. I find this abstraction helpful to convey their general idea, and this is all we shall need in the following discussion (my arguments do not rely on any particular way of spelling the idea out).

What is this idea? It is that our intuitive ascription of dispositional properties relies on assessing a given disposition in a range of *test-cases*. A match will usually not light in a thunderstorm. So, in assessing whether a given match is flammable in the actual world, we will not consider alternative possibilities in which the match is struck in a thunderstorm. A glass will not break when encased in protective layers. So, when assessing whether a glass is fragile in the actual world, we will not consider alternative possibilities in which it is packed in protective layers. What we will instead rely on are cases in which the environmental conditions for manifestation are purged of interferences that we can typically recognize as interferences to the relevant disposition. Moreover, test-cases will be worlds in which the opportunity for exercise is present and the disposition-relevant intrinsic properties of the object have not been changed.

These alternative possibilities are, as it were, laboratory conditions for the manifestation of the relevant dispositions, conditions prepared as *best as possible* for the manifestation of these dispositions. The alternative possibilities which ground the relevant sense in which an object keeps its disposition when masked or finked and loses/does not gain it in cases of mimicking, are then these test-cases:

Test-case (conditional): o has disposition d to φ in response to stimulus S, iff, if o underwent stimulus S and o was in a test-case, o would φ.

We have now made our antecedent significantly more complex. Hence, we are no longer aiming at the wrong sort of modal basis (i.e. close worlds), but at more remote alternative possibilities in which interferences are absent.

I have presented these issues in a counterfactual guise because they are discussed like that in the relevant literature, but it should be obvious that – unlike what some have

claimed (Bonevac & Sosa 2006) – the problem isn't specific to conditional approaches to dispositions. We can express the same idea in a categorical guise:

Test-Case (categorical): o has the disposition d to φ in response to stimulus S iff o φ's in a suitable proportion of test-case-S worlds.

Test-case-S worlds of course are just possible worlds in which the test-case conditions and S obtain. We can now apply these results to the debate about reasons-responsiveness.

2.2. Interferences and Reasons-Responsiveness

Now, as I said, these are the basic resources we need to understand the idea behind complex reasons-responsiveness.

For, as several authors have pointed out[34], Frankfurt cases bear strong similarities to the interference examples discussed above. Frank the Frankfurt agent retains his dispositional property of RR even though the presence of the device makes certain counterfactuals false of him. The device is thus a mask to Frank's RR. It is to Frank's RR capacity what the protective layer is to the glass.

In the same vein, the problem that FSCs pose to RR is that the truth-conditions for the relevant counterfactuals are given by worlds where the device is present, because the device is present in the actual world and the truth-conditions of counterfactuals are given by the set of worlds that is maximally similar to the actual world while still satisfying the specifications of the antecedent. So, in the case of FSCs, the counterfactuals relevant to RR come out as false (Frank would not react differently in the relevant sets of worlds because the device would intervene).

The lesson that authors have drawn from this for the analysis of RR is that just like more complex analyses of dispositions, they cannot rely on simple counterfactuals to characterize RR because of FSCs. Instead, because the problem is caused by conditions in the antecedent which are too unspecific, they also must adopt the strategy of latching onto interference-free worlds by incorporating test-cases into their analysis of reasons-responsiveness. In the orthonomy literature, this is sometimes referred to as *abstracting away* from interfering conditions (Smith 1997, pp. 101-102,). This is because we can think of the process of specifying the relevant test-cases as abstracting away from interfering conditions present in the actual situation. Just as for dispositions more generally, this is either going to involve inserting a test-case clause in the antecedent of our conditionals,

[34] Smith (1997), (2003); Vihvelin (2004), (2013).

or it is going to involve using it to specify the alternative possibilities relevant in the categorical analysis. Either way, it presupposes that we have some idea about what counts as a test-case for reasons-responsiveness. There is significant (and important) room for discussion about what can count as a test-case for RR, and in some sense my arguments in this chapter will explore this room, but I take it that nonetheless a sufficiently clear understanding can be found for now. Test-cases for reasons-responsiveness are cases in which the agent is free from any external distraction or duress which might impact their capacity to pick out and react on the normative reasons of their situation. Further, their environment offers the opportunity to respond to reasons and they have not been tampered with intrinsically. The fact that we hold all (relevant) intrinsic properties of an object constant in test-cases will become important.

The proposal is now to bind RR to more complex counterfactuals such as:

"*If sufficient counter-reasons were present and test-case conditions obtained, Frank would react differently*".

Recall that Frank would not react differently because the device implanted in his brain would force him to react the way he reacts (voting for Trump) even if the balance of reasons was different. In what sense then does Frank remain responsive to reasons? The test-case strategy offers an ingenious answer. Frank would not have reacted differently in close alternative possibilities in which the device interferes. But such conditions are not part of the test-cases for reasons-responsiveness. They are to reasons-responsiveness what thunderstorms are to the disposition of a match to light if struck. In determining whether an object or agent has the relevant disposition, we ca safely abstract away from them and imagine how the object/agent would react in cases in which they are absent. Frank presumably would react differently given a different balance of reasons in those alternative possibilities. And so he remains reasons-responsive in virtue of reacting differently in a sufficient proportion of those test-case worlds.

Notice that interestingly, this strategy also prima facie delivers the correct results for Curious Carmen and Compelled Cody from chapter 1. Carmen of course also has the dispositional property that Frank keeps. But at least if we understand the interferences that the antecedent of the relevant counterfactuals specifies as extrinsic interferences, Cody still would not have reacted differently had sufficient counter-reasons been present. His compulsive desire is what drives him, and so long as we hold this desire constant, Cody's action will be the same in a suitable proportion of worlds.[35]

[35] Can masks be intrinsic so that Cody's desire count as a mask? One counterstrategy to my argument against connecting *RR** with control will rely on the idea that they cannot, that all intrinsic masks to an ability make ascription of the ability impossible. More on this below.

3. Complex Reasons-Responsiveness

The proposal on the table is to understand the difference between compulsive agents and free agents, which include Frankfurt-agents, in terms of a special dispositional property that free agents have. This property is the property of RR to be spelled out in terms of complex modal vocabulary:

> **RR***: An agent is RR* with respect to A iff she responds to sufficient reasons for and against A in a suitable proportion of worlds where such reasons are present and the conditions for test-cases obtain.[36]

In order to avoid cumbersome formulations, I will sometimes call the feature expressed in RR* a *modal profile*. The idea behind grounding orthonomy in RR* is still that free agents have a significantly different modal profile with respect to their actions than compulsive agents. For the sake of brevity, let's call profiles like RR* *open* profiles because they signify the agent's openness and flexibility with respect to reasons. Let's call profiles like that we can expect from the compulsive agent closed profiles because they do not signify the kind of openness and flexibility needed for control.

I consider RR* the best common denominator of the different theories of grounding orthonomy in complex reasons-responsiveness. All of them are committed to *something* like it, although some details will change. I will discuss important differences below. My arguments here don't rely on any of the specifics of *RR** accounts, so relying on the abstract *RR** will render a leaner presentation of my points. In order to avoid the suspicion that I am attacking a straw-man in my streamlined *RR**, let me just briefly point out how two especially well-known complex RR views are related to *RR**.

Writers on reasons-responsiveness seem to be in relative agreement that a modal profile like *RR** must be the ground of control. But they have differed significantly on the *location* of the relevant dispositional property.

The so called New Dispositionalism (Fara 2008, Vivehlin 2004, 2013, perhaps Smith 2003) adopts the position that *agents* are in control in virtue of the modal profile of the agent. This is understood as a novel vindication of the intuition behind standard ability to do otherwise-views in that, according to the New Dispositionalism, it is indeed true that the agent could have done otherwise, albeit in another sense than traditionally assumed. It is true in that the agent does otherwise in a suitable proportion of cases in which the balance of reasons changes, even if they do not have the power to actualize these possibilities.

Take Kadri Vihvelin's complex RR view, according to which what grounds free agency is the ability to choose on the basis of reasons. The relevant ability is given the following characterization by Vihvelin.

> LCA-PROP-Ability: S has the narrow ability at time t to do R in response to the stimulus of S's trying to do R iff, for some intrinsic property B that S has at t, and for some time t' after t, if S were in a test-case at t and S tried to do R and S retained property B until time t', then in a suitable proportion of these cases, S's trying to do R and S's having of B would be an S-complete cause of S's doing R. (Vihvelin 2013, 187)

I will not unpack the many aspects of this idea. What is important is that Vihvelin's variant of the relevant dispositional property is true in virtue of (i) a suitable proportion of possible worlds in which (ii) the intrinsic properties of the agent are held constant and (iii) the circumstances are auspicious in the way they are in test-cases. If the agent does otherwise in these alternative possibilities, they will be in control in the actual scenario. Thus, the new dispositionalism claims, they are in control because they could have done otherwise.

Vihvelin's version differs most strongly from *RR** in that it involves volitional elements as triggers. I presented reasons for thinking that this isn't a plausible way to understand reasons-responsiveness in chapter 1. But whether or not a complex RR view is a *volitionalist* RR view won't matter to my argument.

Fischer and Ravizza (1998) instead adopt the view that agents are in control because their *actual mechanisms* are *RR**. So they think that while it remains true that the agent could not have done otherwise, it's the agent's mechanism that has the relevant dispositional property. A mechanism is a causal system of the agent that brings about the action.

They spell out the relevant dispositional property 'moderate reasons-responsiveness' as follows. A mechanism is moderately reasons-responsive iff it is regularly receptive and weakly reactive to reasons. A mechanism is regularly receptive to reasons iff there is an understandable pattern of alternative possibilities in which there is sufficient reason to do otherwise, the actual mechanism is active, and the agent recognizes that reason. A mechanism is weakly reactive to reasons iff there is at least one alternative possibility in which there is sufficient reason to do otherwise, the actual mechanism is active, and the agent does otherwise (for that reason) (Fischer and Ravizza 1998, 243-244).

Fischer & Ravizza's account deviates from *RR** in that they only require a proportion of worlds for the receptivity component. Reactivity is 'all of a piece' (Fischer and Ravizza 1998, 73), as they say. It enough that there is at least one scenario in which the

mechanism reacts otherwise for it to count as reactive. I argued in the last chapter that this isn't a desirable feature for RR views. Apart from this, the approach utilizes the same abstracting away idea inherent in *RR**, because it relies on holding constant the actually operative *mechanism*, an intrinsic property of the agent responsible for producing action. Thus, mechanisms are reasons-responsive in virtue of alternative possibilities in which extrinsic interferences are absent – alternative possibilities very much like the test-cases in RR* (Fischer & Ravizza 1998, p.38), which vindicates the view that even if the agent could not have done otherwise, the agents mechanism could have.

This essential similarity between the agent-based New Dispositionalism and Fischer & Ravizza's actual-sequence approach – the abstracting away from interferences – finally points to the fact that both approaches are still spelling out control in a way that entails a certain kind of modal dependence claim:

Control$_3$:	An agent S is in control of her φ-ing[37] iff S is RR* with respect to her φ-ing.
Entails:	
Weaker Modal Dependence:	Control over S's φ-ing metaphysically depends on there being a set of *remote* alternative possibilities.

My argument will be that this entailment reveals a deep structural problem for complex reasons-responsiveness. I shall argue that the systems and subsystems that constitute free agency are themselves subject to *specific intrinsic interferences*. These interferences are highly localized, so they won't have much impact on the proportion of worlds in which the agent successfully reacts for reasons. But when they are active, the agent loses control. This will create a mismatch with what control$_3$ predicts. Allow me to express this picture in a metaphor to prime you for what is to come: my argument will be that rational agency isn't a solid block. It is rather like swiss cheese: rational agents are subject to local blindnesses and blockages, irrational agents are subject to local openings. These phenomena, I will suggest, cannot be captured by control$_3$ because of its reliance on the abstracting away strategy.

[37] This formulation is meant to be compatible with views that hold that actions are not things we are in control of, they are themselves 'controllings' as it were – means of exercising agential control.

4. Rational Blind Spots

4.1. The Phenomenon Introduced

In order to appreciate the phenomenon I have in mind, it is worth looking at one of its paradigm manifestations in Greek mythology and then expanding a bit on it.

Think about Achilles.[38] Anointed in the river Styx, Achilles is invulnerable. Invulnerability is a dispositional property. So invulnerability comes with a closed modal profile. The modal profile is such that if a sword blow was directed against Achilles, it would not hurt him and if an arrow was loosened at Achilles, it would not penetrate his skin. So, in a suitable proportion of worlds where conditions such as these obtain, Achilles is attacked in some manner, and no vengeful God is interfering with Achilles' invulnerability, Achilles is not wounded.

Unfortunately, as we all know, this is not entirely true. For Achilles is subject to a weak spot, an Achilles Heel. In the off scenario in which a sword blow or an arrow are aimed at his heel, the attack *will* hurt Achilles.

If Achilles' closed modal profile can be subject to such a phenomenon, then it stands to reason that someone with a flexible modal profile could be subject to it as well. And indeed, we can think of a God who creates a reverse version of Achilles, *Bizarro Achilles*. This man is especially vulnerable to swords and arrows and even lightly thrown punches, except when they are directed at his heel, which is when they will have no effect.

The case of Achilles and Bizarro Achilles is not a special case. We can easily see that it is just an especially salient case of a property that many objects with dispositions exhibit. A glass is fragile. It breaks if dropped from a certain height. But it often happens that, due to the intrinsic structure of the glass, dropping it at exactly the right angle etc. will result in the glass not breaking. A rock is sturdy. But an intrinsic structure of the rock might make it the case that hitting it at the right angle will cause it to break.

Note that on the face of it, these manifestations are not just *flukes*. I am not talking about a rock that breaks in virtue of sheer flukish chance, or a glass that remains intact because of a host of disparate conditions that can only be summed up under the label of freak coincidence. On the contrary, the factors that explain these failures to manifest the disposition under typical conditions when they occur are always the same. Thus, these failures are systematic and repeatable. They are grounded in a property that the object has which will cause the disposition in question to fail to manifest every time the object is in the relevant conditions.

[38] The case was originally made by Manley and Wasserman (2008).

Achilles Heels pose a moderate danger to complex modal analyses of dispositions, because they arguably cannot be abstracted away from. Recall that the strategy I went with in thinking about dispositional properties was to find test-conditions for the disposition, conditions that are ideal for the manifestation of the disposition. Achilles heels cannot be easily subtracted from these ideal conditions, because they manifest in conditions that are otherwise perfect for the manifestation of the disposition. Manley and Wasserman (2008), who brought up such cases in the literature on dispositions, have developed an account based on proportionality designed to deal with Achilles Heels. Since *RR** has a similar clause, it can be seen as an instance of this strategy. So we might wonder why I brought up Achilles Heels at all, as they pose no danger to *RR**. But recall that *RR** is not just an analysis of a dispositional property. It is an analysis of the property that is then used to metaphysically ground and thereby explain the presence of control. My argument will be that Achilles Heels pose a problem for the claim that agents are in control because they have *RR**.

4.2. Rational Blind Spots

In order to get this argument rolling, think about Achilles Heels even more generally. We saw that agents and objects can have weak spots. But I think that it is even more evident that abilities, specifically human abilities and especially rational abilities can be subject to what, in this case, is most appropriately called "blind spots".

Some of our abilities have literal blind spots, like a keen-eyed pilot's ability to spot birds, except when they appear in a particular part of the visual field. Other abilities have more metaphorical blind spots. I have the ability to remain unmoved in the face of Hollywood tear-inducing drama, but that one episode of Orange is the New Black always gets to me. I have the ability to withstand spicy food, but I cannot take that one Indian dish.

In the same vein, our rational abilities clearly often have these kinds of blind spots. We are familiar with these phenomena in our everyday conduct. John is a level-headed, rational man, but don't get him talking about social welfare. Rita is rational, well organised businesswoman, but she regularly makes the wrong business-decisions when her daughter is involved. We can describe all of the above cases within the framework of *RR**. The agents involved in them all have a specific blindness, it seems, to a particular class of reasons. John is not responsive to a certain kind of social-welfare reason (depending on how we spell his case out), Rita is not responsive to a particular class of prudential reasons. Still, of course, in the situations in which these blind spots are triggered, the agents keep their *RR** modal profiles: John can deliberate rationally about politics in general. It is just that a certain type of social welfare topic sends him into an

irrational fit. Rita can make the very same decision without it involving her daughter just fine. Thus, all of these agents would respond to sufficient reasons for and against an action in many situations very similar to the type of situation in which their blind spot precludes them from seeing or acting upon sufficient reasons.[39]

Let me demonstrate this by looking at two more detailed cases, one for the receptivity component and one for the reactivity component of responsiveness to reasons (see chapter 1, section 2 for the two components).

Jealousy

Razvan is a well-tempered, rational man. He is responsive to a range of reasons for and against staying with his spouse related to her fidelity. This involves correctly reading social cues indicative (and not indicative) of infidelity, correctly assessing evidence for and against infidelity, basing his decisions about his relationship on such evidence, and finally perhaps also weighing the severity of breaches of trust should they occur. But Razvan suffers from a blind spot. When he sees his spouse talking to other men in a certain way (smiling at them while tilting their head), Razvan is overcome with great jealousy that prevents him from seeing sufficient reasons pertaining to the fidelity of his spouse. One day, Razvan sees his spouse talking to a man they know (smiling, tilting their head). Overcome with jealousy, Razvan decides to break up with his spouse.

Pushing Buttons

Milena is a calm, contemplated woman. She is very receptive to reasons pertaining to the verbal abuse of people. She can assess the psychological harm it causes, the social disadvantages it has, but she can also see when it is appropriate. And she is very good at translating these reasons into action or omission. Usually, people can be as annoying as they want to be, it will not faze Milena. However, Milena has a problem with one particular guy she knows, David, who just pushes her buttons. When it comes to David, Milena just can't help herself. Even though she sees sufficient reasons not to verbally abuse him, she will just let him have it anyway. Milena will be able to respond to the very same reasons when it comes to any other agent. It's just David and his annoying face who pushes her buttons. One day, David is being annoying again and Milena launches into a particularly vicious tirade against him.[40]

[39] Todd & Tognazzini (2008) present similar cases of blindness.

[40] No notion of 'irresistibility' of these impulses is required for the cases to work.

When Razvan decides to break up with his spouse, he keeps *RR**. This is because Razvan does respond to reasons for and against breaking up with his spouse in a sufficient proportion of test-cases. Razvan's blindness is highly localized – it only triggers in a negligible minority of test-cases. Hence, in most of the test cases in which the situation is just slightly different – cases in which is partner tilts their head and smiles just slightly differently – he will see reasons for and against a break-up. The same is true of Milena. In the overwhelming majority of test-cases, she does react in the way her reasons recommend. This is because most of the test-cases for the capacity to respond to reasons for and against breaking up with someone will feature agents other than David, who fail to trigger Milena's blind spot.

Yet both Milena and Razvan, it seems to me, are agents who fail to be responsive to reasons in their situation. This is after all their failure – that they are blind to reasons or powerless to implement their recommendations. So here, we have agents who seem unresponsive, but have *RR**. This is a problem for *RR**.

Jealousy and Pushing Buttons are cases of blindness or blockage. There are also cases of serendipitous local sight, the triggering of a local openness to rational considerations in an agent who has a non-*RR** modal profile. The most telling version of such a case comes from the literature on so called "rational akrasia" (see Arpaly 2000, Audi 1990). The standard example for rationally acting against one's better judgement here is Huck Finn, who, despite his strong racist views, does not hand his black friend to the slave-traders. Huck Finn is not receptive to moral reasons pertaining to people of colour due to his racism. But he seems locally open to such considerations when they apply to his good unlikely friend. Note again that just like in Achilles' case, these phenomena are hardly flukes. They stem from systematic features of the agent that exhibits them and are repeatable given the very same conditions are held fixed.

However, instead of purely relying on the somewhat complicated case of Huck Finn, I will also make use of the following example:

Frankfurt-Trauma

Due to an old entrenched trauma, Katharina cannot translate reasons for and against verbally abusing people into action properly. Whenever she is faced with a situation in which the relevant reasons become pertinent, she ends up verbally abusing someone. This trauma applies to everyone, except David, with whom Katharina has formed a somewhat mysterious bond. One day David is being very

annoying. Katharina assesses reasons for and against verbally abusing him, and, to her surprise, manages not to let lose at David.[41]

Katharina it seems, is responsive to reasons for and against verbally abusing David in this situation. But her modal profile does not match this judgment. For due to her intrinsic trauma, Katharina will verbally abuse people in an overwhelming majority of cases, thus acting against her best reasons. Notice what has happened here. We admitted about Razvan and Milena from the blind spot cases above that since their blindness is intrinsic to them, it will show up in a negligible minority of the test cases, but because their blindness is so localized, it won't make a difference to whether *RR** is true. With Katharina, the reverse is true. Her trauma is equally intrinsic to her (and perhaps thoroughly entrenched as well), so it cannot be easily abstracted away from. Hence, it will show up in the relevant test-cases, and it will render the result that *RR** does *not* apply to her. But she does have a highly localized opening – a soft spot. If the relevant soft spot is triggered, Katharina becomes intuitively responsive to reasons. But her modal profile does not change. Again, this is a problem for *RR**.

This case is an even more severe problem for the attempt to ground control in *RR** - i.e. for control$_3$. Blind spot cases are cases of lack of responsiveness. And neither are they cases where we would assume that the agent is morally responsible for her actions. This intuition should become even more forceful if you already think that control consists in responding to reasons correctly. For if it is true that in blind spots cases agents are blind to important reasons they have, then they lack the kind of responsiveness that seems so central to RR accounts of control.

The problem is now obvious. While in the cases of Katharina, Razvan, and Milena, the agent's modal profile remains the same, the answer to whether they are in control of their actions changes. So *RR** cannot ground control. In other words, blind spot agents are not in control even though the relevant remote alternative possibilities are present and reverse blind spot agents are in control although the relevant remote alternative possibilities are absent. So blind spot cases falsify the *weaker modal dependence claim* and hence *control$_3$*.

Let me reiterate that what I am not saying in these cases is that there is *no* sense in which the agents in blind spot cases are (for Razvan and Milena) or are not (for Katharina) responsive to reasons. *RR** certainly picks out some sense in which Razvan and Milena keep the capacity to respond to reasons (and Katharina loses it). But it is isn't the sense, clearly, that can ground control over the relevant actions.

[41] Cohen and Handfield (2007) use a very similar case, which they take to be a version of Frankfurt's willing addict, to argue that the abstracting away strategy fails.

I imagine that fans of grounding control in the kind of modal properties spelled out by RR* will have a lot to say about these cases. The rest of this chapter is dedicated to showing how these responses will lead to even more trouble down the line. The narrative here is that advocates of grounding attempts like control$_3$ will think that the occurrence of blind spot cases is a small bug at best and can be eliminated by modifying *RR** in various ways. I think that blind spot cases are a manifestation of a much larger structural problem with the attempt to ground control in alternative possibilities at all. They indicate that there is something fundamentally wrong with the spirit of *RR** and its kin. Reasons-responsiveness in the sense that plausibly grounds control is an actual-sequence notion in a more radical sense, I believe these cases show. The sense in which agents need to be responding to reasons for their actions to count as free requires no alternative possibilities whatsoever.

What can the advocate of *RR** and control$_3$ say against rational blind spot cases then? I will discuss the following lines of defence, taking on what I consider weaker arguments first and the strongest arguments last.

First, they can deny the intuition that Milena and Razvan are not responsible with respect to their actions and the intuition that Katharina is.

Second, they can introduce the idea that we should abstract away from Katharina's trauma after all, making the relevant RR capacity more 'global' (see chapter 1, sect. 4.1.) in the process. These strategies already look unpromising. That is why I will mainly concentrate on the third option, which my reader will perhaps already have anticipated. The third defence strategy is to point out that blind spot cases rely on issues concerning the individuation the relevant RR capacities. The cases can therefore be reinterpreted as showing that we merely need to find a more specific description the relevant RR property in order to get a working version of *RR**. For example, this strategy will claim that Razvan may still have the capacity to respond to reasons for and against breaking up with his partner, but that he lacks the capacity to respond reasons for and against breaking up with his partner *while they smile and tilt their head*. Katharina, it will be claimed, may lack the capacity to respond to reasons for verbally abusing people, but she keeps the capacity to respond to reasons for and against *verbally abusing David* – which is the level of specification relevant to her control and responsibility. I shall argue that this strategy fails because the locality of blind spots outstrips the levels of specification that can – plausibly and logically – be given to agential subcapacities.

5. Counterstrategies of Agent-based Views

One way to reject the significance of my examples at the outset is to deny my intuitive judgments about Razvan and Milena, that is, to claim that they are after all morally responsible for and in control of their actions. We might for example think that Razvan must know about his localized jealousy and should therefore have taken steps to mitigate its effects. In the short term, Razvan should have attempted to withdraw himself from situations that are likely to trigger his jealousy and in the long term, Razvan should have taken steps to get his jealousy under control. He should have gone to behavioural therapy, for example. These considerations express the idea that we are responsible for and, in some sense, in control of compulsive actions if we are in control of whether or not we are in the circumstances that trigger the compulsion. And with this proposal in the background, we can deny that Razvan and Milena were not in control of their actions.

But this strategy depends on changing the exact loci of primary responsibility and control. Razvan and Milena are said to be responsible for the attitudes that produce their uncontrolled actions, and their responsibility for these actions is then understood as a sort of downstream effect. It is fine if we want to say this. But it is neither here nor there in terms of their primary responsibility for and control over their actions – which the strategy admits does not apply in the usual sense. The accounts I am tracking here – accounts that spell out orthonomy in terms of responsiveness to reasons – are attempting to find a primary sense of control, and it is this sense I am claiming is absent in cases of Razvan and Milena. To home in on the important locus of freedom in these cases, we can just stipulate that these agents encounter the circumstances that trigger their blind spots for the first time and that they were previously unaware of their affliction. The question will remain: is Razvan in control of/free with respect to/responsible for breaking up with his partner? Is he responding to reasons? Control_3 says that he is. But surely the intuition is that he isn't.

Additionally, we would still be at a loss as to the status of Katharina in Frankfurt-Trauma. What if she was as negligent as Razvan when it comes to curbing her dispositions? Does she then not count as responsible for her action? Katharina's case affirms that if we hold fixed the locus of primary responsibility and control, it will be significantly harder to deny intuitions.

Thinking about Katharina's case reveals a different, more theory-driven strategy for denying intuitions. The thought goes like this. Katharina seems responsible and in control. So we should treat her intrinsic interference in exactly the same way we treated extrinsic device-like counterfactual interveners. We should abstract away from her interference and consider as relevant test-cases only worlds in which Katharina's trauma is absent (or somehow dormant). There will be questions about how exactly to spell out

RR to achieve this result. That is, there will be questions about what specification of the auspicious circumstances for the manifestation of reasons-responsiveness will render a modal basis which does not contain Katharina's trauma. But I will here grant that it should be possible to find such a specification. Perhaps, for example, we need to think of test-cases as involving idealised circumstances in which Katharina is her best self – untainted by trauma, or at least unflinching in its presence.

Whatever the exact way in which we specify what counts as a test-case for reasons-responsiveness, purging the modal base relevant to Katharina's responsiveness of her trauma will entail that we have to purge the modal bases of Razvan's and Milena's responsiveness of their respective intrinsic interferences. And so we will have to say that they are responsible and in control after all. For in situations in which they are free from their respective blindness, they will do otherwise as well. Note that this strategy is equivalent to claiming that the capacities for responsiveness relevant for responsibility of Razvan, Milena, and Katharina are highly global capacities – capacities purged of almost all features of an agent's actual situation.

This strategy is equally implausible. For it just exacerbates the original problem with Razvan and Milena. Given that their interferences are triggered, Razvan and Milena, as they are in the actual world, are not responsive to reasons. Whether it is true that they retain the capacity to respond to reasons in a very global sense is simply irrelevant to their control and responsibility. When we assume that Razvan and Milena are free and responsible with respect to their actions, we are essentially attributing responsiveness to them based on *vicarious* alternative possibilities. That is, we are imagining what they would have done if they had been not only better agents than they in fact are, but very different agents too. Our blind spots can be parts of who we are – not only to others, but to ourselves as well. Systematic experiences of agential failure are often partially constitutive of identity. Hence, when we ascribe responsiveness to agents based on worlds in which their ideal selves *do* respond, we ascribe responsiveness based on the actions of agents which the actual agent might significantly dissociate from. This is true of Katharina as well. She is responsible and in control, according the strategy currently under investigation, because someone very much unlike her, someone without the kind of psychological burdens that she has had to live with, responds to reasons in a sufficient proportion of idealised cases.

But this should not be what we say about Katharina. What she achieves, she achieves in virtue of what *she as she exists now* can do, in virtue of what she is capable of in her current situation and as the psychologically damaged human that she is. We should not be judged based on the actions of our perfect selves. Moral responsibility and control are about what we can do now, in the situation and body we are in. What our ideal selves

would do is not pertinent to these assessments. It is not pertinent in the same way that it would be irrelevant to tell a person who is drowning due to a muscle cramp that they still *can* swim because their perfect self who is less prone to muscle cramps still does swim in a sufficient proportion of idealised situations (Clarke 2008; Whittle 2010; Franklin 2011).

A third strategy remains. This strategy is to hold that Razvan and Milena do indeed lose the relevant RR capacities – if we focus on the right level of specification of these capacities. The basic idea behind this strategy is that capacities with intrinsic interferences are lost rather than merely masked. But the plausibility of this strategy will depend significantly on whether we focus on *the agent* or one of *their systems* as bearer of reasons-responsiveness. For the remainder of this section, I discuss why the agent-based version doesn't get very far. This will serve as preparation for the ensuing discussion about levels of specification for systems.

The idea behind the third strategy is that an object with an intrinsic opposing disposition (Clarke 2010) seems to be in some way "internally flawed". Imagine for example an agent who is very generous, but also very greedy (both adjectives that ascribe character dispositions, we can assume). Let it be the case that this agent's greed always or most of the time wins over her generosity. There is a sense in which it is untrue then that she is a generous agent. Within the framework I have been working with here, this can be expressed as a "nomic duplicate test"(Choi 2005; 2006; 2012) which is a constraint on what types of interferences can be abstracted away from in making the antecedent of the counterfactuals specific enough. The constraint is that an object has a disposition only if its modal profile also holds for all of its intrinsic duplicates (or "nomic duplicates"). Another way to put this is that we are only allowed to abstract away from all extrinsic interferences (neurosurgeons, storms and the like), while holding constant all of the intrinsic properties of the agent. If that includes intrinsic interferences, then those will feature in the antecedents of the relevant conditionals. Thus, in the case of the generous man, we will have to look at worlds where nothing extrinsically obstructs the man's generosity, but in which he still has the opposing disposition of greed. In those worlds, it will be false that he would give money to the poor if he tried to (or whatever the right counterfactual amounts to here), because they will be worlds where his greed interferes with the success of that action. So with the nomic duplicate test applied, we get the right counterfactual results for the intuition that someone whose greed mostly trumps his generosity is not really generous. In the same way, advocates of the agent-approach argue that agents who have opposing dispositions like jealousy are not really RR* in the

situation, because intrinsic duplicates of them would be equally afflicted by their intrinsic interferences (Fara (2008), Vivehlin (2013) exhibit tendencies to adopt the strategy).[42]

I have two important comments about this strategy.

First, it runs into trouble with the Frankfurt-Trauma case. This is because Katharina's trauma is an intrinsic interference, and so intrinsic duplicates of Katharina will come with the same failure to respond in most circumstances. Hence, the duplicate test will render the result that Katharina *lacks* the relevant capacity to respond to reasons. And this entails that we have to deny the intuition that Katharina is in control of and responsible for her action. This problem is the reverse of the problem with abstracting away from Katharina's trauma. The problem there was that what gets us the right result for Katharina, namely abstracting away from her trauma, gets us the wrong result for Razvan and Milena. The problem now is that what gets us the right result for Razvan and Milena, namely taking intrinsic interferences to erase capacities, gets us the wrong result for Katharina. Complex reasons-responsiveness views are trapped in between those two types of case.

Fara (2008, 585) bites the bullet here, arguing that agents like Katharina are not responsible, contrary to appearances. Basically, his idea is that Katharina is morally responsible and in control with respect to her failure to regulate her addiction only. Because indeed she could not have decided otherwise, she is not responsible for her decision not to verbally abuse David.

I am not sure what this tactic is supposed to achieve. As I already said, we could simply cancel out derivative responsibility via stipulation in blind spot cases. So under this reading, Katharina has done everything in her power to regulate her tendencies, but failed. Still, the basic point about Frankfurt-cases remains: The trauma is not operative in the actual decision, so it seems that this decision/action is controlled. The New Dispositionalist strategy of abstracting away from interveners is so attractive precisely because it can account for this intuition so elegantly. The advantage of the New Dispositionalist answer to Frankfurt-cases is that it does not have to claim that the agent is not responsible despite appearances because they lack the ability to do otherwise,

[42] Kadri Vihvelin is a special case. Sometimes she tends towards the same strategy as Smith and Fara, arguing for example, that phobia and psychosis rob agents of global abilities (Vihvelin 2013, p.203). In this case, she also falls victim to my examples. At other places, she admits that global abilities (narrow abilities in her terminology) are not all that free will amounts to. But of course, this concession is in tension with her treatment of FSCs (and probably her treatment of intrinsic FSCs, had she ever responded to their threat). So she always accompanies it with saying that people nevertheless keep the free will "they think they have" by keeping global abilities (Vihvelin 2013, p. 169). But what is the free will "we think we have"? If I am right, the free will people think they have is the control property, that in virtue of which they are in control of particular decisions. As my examples show, this property is not a global ability but something like a local ability. The core of the common-sense free will concept is not just having the intrinsic capability, it is having that capability and its being up to oneself to exercise it (or something along those lines). Thus, even if Vihvelin admits of cases like that of Razvan and Milena in her project, she fails to assign them the right significance in the dialectic. What these examples show that control, most fundamentally, is not a global ability.

but that they keep this ability and this is why we have the intuition that they are responsible in the first place. Fara's response loses this distinct advantage, thus abandoning one of the most compelling reasons to adopt a complex reasons-responsiveness view in the first place.

More importantly however, it is unclear whether the agent-based version of the 'intrinsic interference = erasure' strategy even gets off the ground. For notice what has happened here. The example of the generous man who is greedy works only if the man's greed trumps his generosity most of the time. If there are enough situations where his generosity wins, the proportionality clause renders a positive result for the disposition to be generous. If the man gave to the poor in a sufficient proportion of cases where he is faced with poverty, even though in some of them he would greedily turn away, then it is still true that he is generous.

The point of blind spots is exactly that they escape the proportionality clause because they only trigger in very specific conditions. So, if we think about reasons relevant to Razvan's decision to break up with his partner, then the blind spot will only trigger in a very specific subclass of situations where Razvan needs to be attentive to fidelity-reasons. In a suitable proportion of these cases, Razvan is responsive to such reasons. Only when his partner smiles at an interlocutor in a certain way, Razvan's blind spot will trigger. This is why the threat blind spots pose to a modal analysis of dispositions disappears once we adopt a proportionality clause. They trigger in only a fraction of the worlds relevant to the truth of the relevant counterfactuals. But the point here is again that adopting a proportionality clause is exactly what makes blind spots so dangerous once you want to ground control in the modal profiles. That is, blind spots are dangerous for the connection between control and modal profiles because the proportionality clause serves to preserve the modal profile in cases of failure that are nevertheless part of the relevant class of possible worlds. The problem here is that the modal profile of agents who apparently lose control should not stay the same in cases of a triggered blind spot.

This failure to account for blind spots by theories that focus on the agent as a bearer of the RR dispositional property exposes what perhaps was already obvious: the force of blind spot cases prima facie relies on thinking the relevant capacity in Razvan's circumstances is the *capacity to respond to reasons for and against breaking up with his partner*. This is the capacity that Razvan keeps in the *RR** sense even though he has a blind spot. But maybe all my cases show is that this capacity is too general to be relevant. Perhaps what is really important is his *capacity to respond to reasons for and against breaking up with a partner who is smiling and tilting their head*. Razvan lacks this capacity in the *RR** sense because most alternative possibilities relevant to this capacity contain

instances in which Razvan's blind spot is triggered. The remainder of this chapter discusses this type of response.

6. Mechanism-Based Approaches, Blind Spots, and Individuation Trouble

Instead of focusing on the agent, Fischer and Ravizza (1998), (as well as, in my interpretation, Smith 1997, 2003, 2009) propose to focus on the agent's *actually operative mechanism*. This move correlates to a natural reaction one might have with respect to blind spot cases. We might think that, while it is true that *Razvan* could have reacted otherwise to sufficient counter-reasons, Razvan was being compelled by his jealousy, and his jealousy, that which actually brought about his action, was *not RR**. So under this proposal, we need to look at the modal profiles of mechanisms rather than agents in order to determine whether the agent's actual action was free. The proposal is of course closer to actual-sequence intuitions in that it assumes the bearer of RR must be the actually involved in the production of action.

What are mechanisms though? Fischer and Ravizza never gave an answer to the question how we individuate the agent's actually operative mechanism unfortunately, apart from the remark that it is whatever actually brings the action about – an unhelpful answer at best.[43] Skipping a lengthy discussion about the individuation criteria of mechanisms, I think it is safe to assume that mechanisms do not comprise the entire causal history of the action. So a mechanism will not include temporally and agentially distant causes like the Big Bang. We can add, with Smith (2003, 25), that part of the relevant history of the action are the intrinsic properties of the agent that are causally active when she is exercising her reasons-responsiveness.

Before, I suggested thinking about an agent as a system of complex interrelated dispositions. The move from agents to mechanisms in this framework amounts to a move from the general system to the subsystem that is explanatorily most relevant to the agent's actual action. Agent-based proposals track the agent through a modal subspace and look at their reaction. Mechanism-based proposal track a subsystem (or several of them) of the agent through a modal subspace is to look at the outcomes that system produces.

Note how this prima facie gets rid of cases like that of jealous Razvan, Milena, and Katharina. Relevant to the assessment of Razvan's control, according to the current proposal, is his *capacity to respond to reasons for breaking up with someone who smiles and tilts their head in a specific way*. The modal base for this capacity consists of

[43] Issues with individuation of mechanisms in Fischer and Ravizza (1998) (which are a special case of the general problem discussed here) are brought up in McKenna (2001), Watson (2004), 298, Ginet (2006), 234–235.

situations in which someone displays the relevant behaviour and Razvan, as he is intrinsically, is faced with the decision whether to break up with them or not. And in an overwhelming majority of these worlds, Razvan cannot see reasons. For worlds in which someone smiles and tilts their head are worlds where the blind spot is triggered. Hence, Razvan lacks the capacity to respond to reasons for breaking up with smiling-while-tilting their head people (the same narrative holds for Milena). Consider next Katharina. She seems to be in control with respect to her omission to hit David. According to the current proposal, she has the *capacity to respond to reasons for and against verbally ab using David*, but she lacks the capacity to respond to reasons for and against verbally abusing everyone else. If we track Katharina's system under this specification – the capacity to respond to reasons for and against verbally abusing David –, it will turn out that we get the correct RR results. For this system *does* produce the fitting responses in a suitable proportion of worlds where the relevant situations obtain – situations in which Katharina decides whether or not to verbally abuse *David*. We can understand this strategy as a more specific duplicate test: We specify the actually operative system, and we look at whether its intrinsic duplicates will produce the correct actions in a sufficient proportion of idealised cases. The systems of Razvan and Milena don't. Katharina's system does.

I have two basic responses to this strategy. It is (i) implausible and downright incoherent to assume such a fine-grained individuation of systems, but even if it is granted that systems can be picked out with such a high degree of specificity, there will be (ii) new blind spot/weak spot cases for those systems.

Let me anticipate the larger point that should be emerging from this discussion. In order to adhere to the idea that RR can be spelled out via modal profiles, complex RR approaches need to invent a particular agential metaphysics. This metaphysics follows an idea of purity: some systems have open modal profiles with respect to reasons-responsiveness. These systems produce free actions. Other systems have closed modal profiles with respect to reasons-responsiveness. These systems produce unfree actions. But my blind spot cases indicate a different picture: agential systems aren't so clear cut. The very same systems may, under certain conditions, produce unfree actions even if they usually produce free actions. In these cases, the system keeps its open modal profile, but the action is nevertheless unresponsive and unfree. What matters is then how the action is actually produced.

(i) System Individuation

On what basis can we assume that the systems that produce Razvan's and Milena's actions should be picked out under such highly specified descriptions as 'capacity to

respond to reasons for and against verbally abusing David' or 'capacity to respond to reasons for and against breaking up with people smiling and tilting their heads'?

I think there are three interrelated narratives here that (would seem to) support such a point. The first is that given the way in which Razvan and Milena's actions are described, they count as compulsive. After all, in some sense both Milena and Razvan cannot help but perform these actions. Hence, their actions must be brought about by, as it were, compulsive systems, systems that do not exhibit the right modal features. The second narrative is that Razvan especially is described as jealous. And so the impression is that what causes his action is his jealousy, not a reasons-responsive system. The third narrative is that we should pick as the operative system the system which is proximally causally responsible for the relevant action. And what causes Razvan's action, again, is his jealousy.

Let me address the first two narratives first. Jealousy, we might think, is the paradigm of the type of emotion that just overwhelms us, the stereotype of a compulsive force. Compulsive action, we might further assume, is action not performed based on responsiveness to reasons, but on brute causal force alone. So what actually explains Razvan's action is his brute compulsion under jealousy. The system from which he acts is obviously not responsive to reasons. In fact, this is how RR theorists have often, implicitly or explicitly, thought about these cases. Michael Smith, for example, suggests that in order to test whether an agent's failure to act in accordance with reasons is explained by his failure to exercise his rational capacities, rather than "self-hatred", we should look at her capacity and hold her emotions (including self-hatred) constant (See Smith 2003, p.30). If the resulting modal profile is closed, the actual explanation of the failure to act is self-hatred.

However, if we keep the dialectical situation in mind, this way of picking out systems is uninformatively circular. We started out with the distinction between free and compulsive agents. RR theorists account for this distinction by pointing out that free and compulsive agents have differing modal profiles. I, in turn, pointed out that there are cases where the agent is intuitively not in control, yet her modal profile suggests otherwise (and vice versa). The response to this objection is that her actually operative system *does* have the right modal profile. In determining what exactly the right system is, the RR theorist then again appeals to the fact that the modal profile of the system is a closed profile. But the point of my examples is exactly that there are systems with closed profiles that nevertheless produce controlled actions and systems with open profiles that produce compulsive actions. So in appealing to the modal profile of the system in determining the identity of the system in the first place, the RR theorist is blatantly begging the question against my examples. Apply this to Razvan's case. In my opinion,

the actually operative system in Razvan's case is system with the intrinsic bind spot that often responds to fidelity reasons. But the advocate of the mechanism-based approach counters that the actually operative system is Razvan's jealousy. Why? Because it is the system with the correct closed modal profile to match Razvan's lack of control, the advocate of the mechanism-based approach replies. But why should we accept this individuation criterion unless we were antecedently convinced that all compulsive, unfree action corresponds to closed modal profiles? That this is so is exactly the assumption that Razvan's case is supposed to disprove. So falling back on modal profiles to substantiate the restriction to jealousy alone is begging the question against blind spot cases.

Second, reflection upon the correct description of blind spot cases lends support to the assumption that a larger system is involved in explaining Razvan's action. Here is how I imagine the scene of jealous Razvan to play out. Razvan watches his boyfriend/girlfriend talking to another man, smiling. Pondering what to make of this scene, he tries to respond to reasons to alter (or not alter) their relationship related to his partner's fidelity (which for him, is a trumping reason for deciding to break up). He looks and looks, but the more he looks, the more evident his spouse's betrayal seems to become. Razvan is looking straight at the evidence for his spouse's fidelity, but he cannot see it, that is, he cannot form beliefs about reasons for her fidelity due to his jealousy. We do not need to imagine that Razvan is overcome with emotion. Although this is how some types of jealousy work, many other types are calm, anaemic emotions, that obscure the vision of reasons from the background without becoming especially salient to the agent. Razvan goes on to weigh his reasons for and against his spouse's fidelity, becomes convinced that she is having an affair and consequently decides to break up with his spouse. There is no reason to think that this scene is falsely described by saying that the actually operative system was the system that is causally responsible for responding to reasons for and against breaking up with someone. To reiterate, this system has an open modal profile, as Razvan is usually responsive to reasons of fidelity. Thus, if we want to use Smith's terms for example, the case looks like one where Razvan's failure to decide correctly is explained by his actually operative system having a flaw. The full explanation of his failure is not the flaw alone, but the fact that his usually responsive system had that flaw.

To be clear, of course the intuition that if Razvan and Milena are not in control, then they cannot be responsive to reasons in their situation, is correct. There is a sense in which Razvan und Milena are evidently unresponsive to reasons. The point here is that it does not seem to be the modal sense spelled out by *RR**-like approaches, at least if we individuate systems in a plausible manner.

There remains the third narrative: the idea that what causes Razvan's and Milena's actions is the relevant system and the causally potent system in Razvan's case, for example, is his jealousy. The causal strategy is inherent in Lewis idea discussed above that test-cases are worlds in which the relevant intrinsic properties of an object are the proximate cause of an outcome and in Fischer and Ravizza's idea of mechanisms. You can perhaps already see that this criterion is neither here nor there. For the point currently under discussion is how we should think about this system. Is it the same system that is active when Razvan successfully sees reasons for and against breaking up with his partner, or is it a different, unrelated system?

The central problem for the idea of identifying the freedom-relevant system as the causally operative one is that systems may cause actions in more than one way. When the motor of a car makes that car stall because it is broken, then we still say that *the motor* made the car stall, not whatever it was that caused the motor to break. We can see this better by thinking about cases in which the motor brings about a typical manifestation behaviour in the car – driving – but not in the right way. For example, the motor might be so broken that when I hit the gas pedal, it explodes, thereby propelling my car forward a good bit. My motor then did, superficially at least, what it tends to do when I press the gas pedal. But it did it in the wrong way. This is a case of causal deviance, in which a causal system brings about its typical manifestation type under auspicious circumstances, and yet we would not count the outcome as a manifestation of that system. I will look at cases of deviance throughout this thesis (see especially chapter 6, section 5). Here, what is important is that it is understood in these cases that the *very same system* can either act as the right type of proximate cause or the wrong type of proximate cause.[44]

This is why appealing to the proximate causes of Razvan's and Milena's actions is neither here nor there – because we can think of their less finely individuated systems ('the capacity to recognize reasons for and against breaking up with a partner') as causing outcomes in deviant ways.

The phenomenon of causal deviance is of central significance to this chapter more generally. For in undermining any attempt to pick out the right system in terms of its causal role, it directly establishes that the possession of a capacity (no matter what its description is) cannot be enough to ground control. This is because in cases of deviance a dispositional system causes its typical outcome and yet that outcome does not count as a manifestation of that system. This principle holds for responsiveness to reasons as well. Take for example Daniele (a cousin of Razvan). Daniele too is highly jealous in

[44] Notably, Smith (2009) develops his orthonomy view from the insight that deviance requires the notion of a capacity. It is however not entirely clear whether Smith is advocating an exercise account or a possession account.

situation in which his partner smiles and tilts their head. But just as in the case of the exploding motor, Daniele will always decide to stay with his partner when his blind spot is activated. When Daniele decides to stay with his partner, he deserves as little credit for his decision as Razvan does for his. For Daniele too is not in control of his decision. But Daniele keeps his capacities to respond to reasons (be they global or local). The issue with him is not what capacities he possesses, but what capacities he exercises/fails to exercise. More precisely: In cases of deviance, the agent keeps their relevant capacity, and the system which grounds that capacity is causally active. But because the systems brings about the relevant outcome deviantly, that outcome does not count as a manifestation/exercise of the capacity linked to that system.

Hence, it cannot be enough in general for an agent to *possess* a capacity to respond to reasons for their action to be free. They must also *exercise* these capacities in order to be free with respect to the relevant actions. Deviance cases undermine what I called in chapter 1 Possession Views of Orthonomy. They support Exercise Views of Orthonomy.

None of the three stories we can tell about why we should individuate system in a more fine-grained way is plausible. There is additionally also very good reasons not to individuate systems in that way.

Think first about the plausible generally accepted condition that systems (capacities, dispositions) are individuated in part through the type of manifestation they tend to bring about.[45] Take for example the ability to sing David Bowie songs. We can imagine an agent who will produce David Bowie songs in a sufficient proportion of test-cases in which they intend to sing the relevant song, except in cases in which they intend to sing *Life on Mars*. Due to some emotional scar, whenever they intend to sing *Life on Mars*, they tear up and choke up. Now, if we are assessing the ability to sing David Bowie songs, the inability to sing *Life on Mars* will seem like a blind spot. That is, in circumstances in which our protagonist intends to sing *Life on Mars*, they will keep an open modal profile with respect to the ability to sing David Bowie songs. They sing David Bowie songs in a sufficient proportion of test-cases in which they intend to sing David Bowie songs. If the possession of *this* ability is relevant to, say, assessing whether we should blame the failure to sing a David Bowie song on the agent, then we will get unintuitive results. For the modal profile tells us that our protagonist keeps this ability, while our intuition tells us that they should not be blamed for their failure.

However, akin to the specification strategy for *RR**, we can very plausibly hold that the ability relevant to our blame/praise is the ability to sing *Life on Mars* – a more fine-grained system. This system of course has a closed modal profile. Our protagonist fails to sing

[45] See for example Vetter (2014).

Life on Mars in all or most worlds in which they intend to sing *Life on Mars*. So it looks like the specification strategy is very successful for this example.

But now consider an example in which the blind spot of our protagonist does not line up so neatly with an action type. Assume that the emotional scars of our protagonist are even more complicated. Instead of intending to sing a particular song, the emotional baggage will make the agent choke up when they intend to sing a David Bowie song when their ex-lover is present. Is the ability relevant to blame in this case the *ability to sing David Bowe songs when your ex-lover is present*?

The original example works because singing *Life on Mars* offers a more specific description of the action-type in question. But while singing *Life on Mars* is a type of action, singing *Life on Mars* while your ex-lover is present is not. Singing *Life on Mars* while your ex-lover is present is performing the action type: *Singing Life on Mars,* in circumstances in which your ex-lover is present. That is why it makes little sense to specify the system in the suggested way. Abilities are in part individuated by the type of thing they produce. Singing David Bowie songs and singing David Bowie songs while your ex-lover is present is the same type of thing, brought about in two types of circumstances.

This problem is exacerbated in the domain of reasons-responsiveness. The reasons agents are responding to for the practical domain are *reasons for action*. This is why it makes sense to describe Razvan as possessing the capacity to respond to reasons for and against breaking up with his partner. Breaking up with someone is a type of action, to which often enough relevantly similar considerations apply. But breaking up with someone in circumstances in which they talked to someone while smiling and tilting their head is *not* a separate type of action for which separate types of considerations are relevant. It is just breaking up with someone in different circumstances. In the same vein, verbally abusing someone is an action type. Verbally abusing *David* is performing an action of that type against David. Hence, it does not correspond to our intuitive ways of ascribing agentive abilities to specify them infinitely. Capacities to respond to reasons have intuitive stopping points for how fine-grained we can think about them. And the description: 'capacity to respond to reasons for and against breaking up with someone' is already pushing the limits.[46]

That was my first point. It depends on the plausibility of not individuating action-types in a highly fine-grained manner. Perhaps this point will be found unconvincing. Let me therefore add my second point.

[46] These considerations are controversial because they assume the environmental circumstances to be outside of the scope of the ability description (Fisher 2013 calls this a 'two-parameter analysis).

Think about the resources that RR theorists use in their theory. These resources just straightforwardly entail the possibility of blind spots in the way I am envisaging them.

If we think about how RR theorists conceive of the disposition they take to be relevant to control, it is no surprise that cases such as Razvan's are possible. Recall what it means that a system has a modal profile. In the case of an open modal profile, it means that *this system* would produce different actions in slightly different circumstances. This means that RR theorists are beholden to the assumption that systems are multitrack[47]: the very same structure of causally interrelated properties may bring about a whole range of manifestations. All I am adding to this assumption is that, imperfect agents that we are, our systems too have intrinsic counter-dispositions. And this is just the long and complicated expression of the thought that I started out with: That our *abilities and dispositions* have bind spots. It is not only *the agent* who has these blind spots, but the agent's subsystems that ground her abilities. These blind spots block the typical manifestation of a system's causal output under very localized conditions, so that they do not change the modal profile of the system. But if they are triggered, they rob the agent of control. Unless the RR theorists give us good reason to assume that subsystems cannot be intrinsically opposed, their adherence to a multitrack view will leave open the logical possibility of blind spots. My cases show this possibility to be a very real and actual.

In a similar vein, think about the cases of severe addicts that I excluded via the proportionality clause again. Even severe addicts, RR theorists admit, respond to dramatically sufficient counterreasons. The most hopeless heroin addict would not go for the shot if the house around him was on fire. In response, RR theorists can introduce the idea that only a suitable proportion of counterfactuals must be true of the agent/mechanism in order to be *RR**. But think about what they are thereby admitting: They are conceding that there are possible scenarios in which an agent acts *by way of the very same mechanism* that in the actual scenario counts as compulsive and in so acting does respond to a dramatically sufficient counter reason. All I am asking is what happens if one of these possible scenarios becomes actual. According to the modal profile of the mechanism in question, we should count the addict who leaves his house instead of taking the drug as not RR and therefore lacking control. But we just admitted that in a way, she does react to the reason of his house being on fire and intuition certainly point towards his being in control.

The open causal pathways or opposing powers that function as blind spots may not even be conceived of as some wayward, alien element, even though I think often they are

[47] The terminology comes from Ryle (1949), 43-45, but can be found in contemporary literature as, for example in Heil (2003). Vetter (2013) uses the terminology to denote dispositions that 'can't be characterized by a single conditional', which I don't think is clearly equivalent to the Rylean usage.

(think of phobias, traumas etc.). Some systems may be varyingly responsive and unresponsive because it was evolutionarily advantageous for them to develop this way in human phylogenesis. Think about an agent who acts out of fear, for example. Sometimes we conceive of fear as a compulsive force, which propels us forward irrespective of the reasons there are for doing what we do. Other times, we think of fear as a mechanism that is very responsive to environmental factors that pose threats to us. So a man running out of a house out of fear may either act compulsively or responsively, depending on how we describe the function that fear has in his actions. Importantly, it seems natural to assume that his fear system remains the same in both scenarios. It is just that this system has manifestations that we intuitively count as compulsive and manifestations that we intuitively count as responsive. The same thought is expressed by Michael Mckenna when he suggests that it is a "deep problem" for mechanism-based views that

> […] any complex system will have "subsystems" that are designed to function precisely by shutting down or by permitting other systems to override in some contexts but not others. (McKenna 2013, p.164)

He illustrates:

> Reflection on complex mechanisms like automobiles or computers, things that are built up out of smaller mechanisms, helps to illustrate the point. Suppose that one of a computer's programs runs unimpeded, but the computer is designed to divert that program and prioritize other operations if the system is under stress, or uploading new software, or about to lose power, or what have you. The same applies to an automobile. A car will allow things like unimpeded gas flow through the fuel injection system unless another part of the system recognizes problems with the fuel mix or something of the sort. If we were to evaluate the degree of sensitivity of such a system by "holding fixed the actually operative submechanism" we'd hamstring the system for a variety of conditions to which, without holding these fixed, the larger system as a whole would be able to respond quite easily. (McKenna ibid)

This is making exactly the point I am trying to make about how blind spots work. I would just like to emphasise, contra Mckenna's way of putting the thought, that blind spots can occur in even very finely individuated systems. As I said above, since blind spots can be very local, we need not think of the "larger subsystem" as akin to the agent, as Mckenna thinks.

These points to my mind thoroughly undermine the strategy of specifying the freedom-relevant systems in blind spot cases. But I have encountered one last way to try to avoid my cases, which is to argue that it is at least logically possible to specify subsystems

indefinitely. If this statement is true, then there will at least be a logical exit strategy for *RR**-like approaches. But I do not think it is possible. This is because blind spots outstrip *any* level of specification of subsystems.

(ii) Infinite Specification

Let us grant what I denied in (i), namely that it makes sense to describe Razvan's freedom-relevant system as the capacity to respond to reasons for and against breaking up with people in circumstances in which they smile and tilt their head. Does this solve the problems with blind spots? It doesn't. For I can tell another story about Razvan-2, who responds to reasons in circumstances in which his partner smiles at people just fine, but whose blind jealousy is only triggered in cases in which his partner is smiling while tilting their head to someone with a moustache. Razvan-2 responds to reasons for and against breaking up with someone in sufficient proportion of test-cases, but when his partner smiles and tilts their head to someone with a moustache, he can't see reasons properly. What if these blind spot circumstances become actual? Again, I hope our intuition will be that Razvan is not (or less than fully) responsible, because he lacks control over breaking up with his partner. But he keeps an open modal profile with respect to his RR capacity to respond to reasons for and against breaking up with people in circumstances in which they smile and tilt their head.

My opponent can of course insist, with even less intuitive plausibility, that now I have shifted the relevant system again, that the freedom-relevant system now is the capacity to respond to reasons for and against breaking up with someone in circumstances in which they smile and tilt their head in conversation with someone wearing a moustache. If the prior specification was implausible, this one is surely ridiculous. But it is possible, nonetheless.

As you can see, the dialectical situation now looks like it is about to hit a tiresome vortex, in which ever more specific capacity descriptions will be countered by ever more localized blind spots. To avoid a methodologically unrewarding exchange like this, let me attempt to present my blind spot argument in a way that shows how any level of specification of RR capacities will face potential counterexamples. The full philosophical significance of this generalised argument will probably not be clear from at this stage of the thesis. It is to be read in tandem with the results of chapter 3 and 7.

The generalised blind-spot argument is simple. Imagine what still needs to be true of even the most specific RR-capacity if the *RR**-like approach is to go through. Call this the capacity$_{\Omega}$ to respond to reasons, in order to eschew the exhausting full specification. Now we need to ask what needs to be true of the capacity$_{\Omega}$ if it is to count as a capacity

that still adheres to the complex reasons-responsiveness model. The model holds that agents in Frankfurt-cases are still responsive to reasons even though they don't respond different to different reasons in close worlds, they do respond differently to different reasons in test-cases – remote worlds purged of actual interferences. Hence two things need to be true of the capacity$_\Omega$. It (i) must still have at least *one* alternative manifestation type. That is, it must still be a capacity such that both the agent φ-ing (in response to the balance of reasons supporting φ-ing) and ψ-ing (in response to the balance of reasons supporting ψ-ing). It must (ii) be a capacity that is ascribed on the basis of remote worlds in which interferences that count as extrinsic are absent (and opportunity for manifestation exists, and no intrinsic changes to the agent were made).

These two ingredients, which are essential to the idea of complex reasons-responsiveness, are enough to generate a problem. For consider a highly localized intrinsic blind spot which cuts across φ-ing and ψ-ing situations such that the agent fails to φ or fails to ψ in only a very small subset of these situations. When this blind spot is active, the agent will keep the capacity$_\Omega$ to respond to reasons, but they won't be responsive in the sense that grounds control. We can also create the corresponding Frankfurt-case, in which some highly localized intrinsic trauma will prevent an agent from φ-ing or ψ-ing in a large number of test-cases, but they still respond to reasons for φ-ing or ψ-ing in the actual world, because the trauma isn't triggered there. Such an agent is still responsible for what they are actually doing, but they will fail to respond to reasons on the same basis in most test-cases.

The problem that blind spot cases uncover is therefore not a problem of making the freedom-relevant systems more specific. It is rather the more fundamental problem that for each modal profile of a system, we can find cases in which the system produces actions which do not fit that profile. This is a problem with spelling out reasons-responsiveness in terms of alternative possibilities at all.

7. The Larger Point of Blind Spot Cases

In conclusion, I think there is very good reason to assume that agents and systems have blind spots and that this falsifies *control*$_3$ and the *weaker modal dependence claim*.

For this last section, I want to try and indicate what I think the general significance of my discussion so far is. In order to do that, let me point out another class of cases that I tried to avoid mentioning until now. We have heard of Frankfurt agents like Frank, compulsive agents like Cody, normal agents like Carmen and blind spot agents like Razvan. In addition to cases like this, *akratic actions* place an important constraint on theories of

control. In the wake of Davidson's treatment of akrasia[48], acting akratically is often defined as acting intentionally and freely against one's unconditional (as opposed to Davidson's conditional best judgment) judgement of what is the overall best thing to do. This definition is a problem for RR theories in particular, because on most accounts of practical reasoning, acting against one's better judgment entails a failure in reasons-responsiveness. So akratic action is free, but not reasons-responsive, it would appear. Akratic actions hence occupy an uncomfortable middle ground between compulsion and free action.

The most interesting general point about RR theories that my prolonged discussion of blind spots reveals is that they are premised on a metaphysics of agency that is shaped to fit the distinction between compulsion and akrasia. The systems of our agency, on this metaphysics, are sharply divided between systems that lead to compulsive actions if operative and systems that lead to free actions if operative. This way of dividing up agency ensures that akratic actions come out as minor failures of an otherwise responsive system, while compulsive actions come out as unfree. In this division, no logical room is left for systems that are highly malleable in terms of what kind of action (free or unfree) they produce. Since blind spots are, as I argued, best described within a framework that allows for such systems, there is tremendous pressure for RR theorists to either understand them as compulsive actions (jealousy is a purely compulsive system) or akratic actions (Razvan could have pulled himself together and overcome his jealousy). Let me point out this tendency to erase logical space for blind spots in order to accommodate the divide between akrasia and compulsion in some RR authors.

The strategy is most obvious in the case of Smith whose project is explicitly motivated by distinguishing akrasia from compulsion. According to Smith, what distinguishes akratic agents from compulsive agents is that the actually operative system of the former is *RR** while the actually operative system of the latter is not. In Smith's terminology, both acratic and compulsive actions are produced by a desire to perform them (and a belief that it would be best to desire something else). But while akratic actions are explained by the agent's failure to exercise a rational capacity that she has (a capacity to desire what coherence demands of her), compulsive actions are explained by a desire that the agent cannot resist in the sense of lacking a rational capacity (to desire what coherence demands of her). Blind spot actions, in this terminology, would be actions explained by the agent's failure to exercise her rational capacity that are nevertheless not free. This is a combination of verdicts that Smith's account will not admit: Either an action is explained by the agent's failure to exercise a capacity she has, in which case it is free. Or

[48] Davidson (1970c)

the agent does not have that rational capacity, in which case her action is explained by some irresistible impulse and is not free. There is no middle ground.

Fischer and Ravizza's approach is also shaped by their response to akrasia. They imagine a case of a severely akratic agent who takes a "non-addictive drug" and would refrain from taking it only in one special situation. Since they assume that akratic action is free (and responsible), they assume that the actually operative system in such a case must be *RR**. Consequently, they limit their version of *RR** in an important way. They claim that while the actually operative system (mechanism, as they say) must be receptive to reasons in a suitable proportion of possible cases, it suffices that it reacts to those reasons in only one case. This is because "reactivity is all of a piece": If a mechanism does react in one scenario, it *can* react in all of them. As I pointed out in chapter 1, this is clearly too weak. Any reasonable account of compulsion will admit that even a compulsive mechanism reacts to the occasional dramatic sufficient reason. Nevertheless, such a mechanism will usually produce unresponsive actions. The only way that Fischer and Ravizza can avoid the collapse of the distinction between akratic and compulsive actions is then to adjust their metaphysics of mechanisms: If we imagine the compulsive agent acting for a dramatic counter-reason, we should actually imagine her to be acting for a different mechanism, a *RR** mechanism. On the other hand, if we imagine the akratic agent reacting differently in only one scenario, we should imagine her acting by way of the very same mechanism active in cases where she fails to act. The possibility for cases of agents who are unfree but act from responsive mechanisms naturally flows from the cases the Fischer and Ravizza consider. But they nonetheless forcefully restrict their metaphysics in order to exclude that possibility.

An implausible agential metaphysics is one component of why *control*$_3$ runs into problems with blind spot cases. The other component consists in understanding reasons-responsiveness purely in terms of alternative possibilities. Evidently, the kind of responsiveness to reasons that is lacking in Razvan and Milena's case cannot be spelled in terms of what happens in a range of alternative possibilities – no matter how specifically we try to narrow down the relevant modal base. Razvan and Milena remain responsive in the *RR** sense. But it is not the sense in virtue of which agents are free when they act (because Razvan and Milena are not free when they act). Rather, as I indicated in section 6, Razvan and Milena are not in control of their actions because they fail to *exercise* their responsiveness to reasons for the action they are actually about to perform. What grounds control is then not a *RR** type dispositional property.

The result that *control*$_3$ is false because the entailed *weaker modal dependence claim* is false is a problem for New Dispositionalist accounts that seek to rejuvenate the ability to do otherwise with their analysis of complex reasons-responsiveness, because it shows

that even if traditional Frankfurt-cases can be overcome with their analysis, blind spot Frankfurt-cases like Frankfurt Trauma cannot be treated in this way. Neuroscientists and devices are easily abstracted away from. But identity-constituting traumata are intrinsic and deeply entrenched, which makes them hard or indeed impossible to deal with for the abstracting away strategy.

However, the falsity of *control*$_3$ is additionally strangely puzzling for those views that seek to accommodate Frankfurt's original insight, which I expressed in the last chapter as the claim that control is exclusively grounded in features of the actual sequence. For their account are explicitly built to be *actual-sequence* accounts. And yet they apparently fail to capture in what way control ought to be grounded in actual-sequence features. This puzzle will be addressed in the next chapter.

Let me merely point to here the path I see for an account that can overcome the problems discussed in this chapter. I have tried to argue that we should embrace cases where our judgements about control (and responsibility) run contrary to what the dispositional properties of the actually operative system, most plausibly individuated, suggest. What I have tried to show in this last section is that we can only really admit blind spot cases as genuine if we let go of the metaphysics of agency that is explicitly or implicitly assumed by RR theorists, the kind of metaphysics designed to keep apart compulsion from akrasia on the level of systems.

What I am rejecting in committing to blind spot cases is then this package: The assumption that control over actions is grounded in a dispositional property plus a way of individuating that property on the basis of whether or not the action in question is compulsive or akratic.

Blind spot cases show that systems may keep the very same modal profile while fluctuating in whether actions produced by them remain under our control. Razvan's system that contains the jealousy produces controlled action most of the time, but sometimes produces uncontrolled, compulsive-looking actions. Even Cody's system, which produces compulsive action almost all of the time is sometimes responsive. The actions so produced seem free. A project starting from this kind of metaphysics does not necessarily rule out that all akratic actions are free. But it is not limited by the idea that they must be either. More importantly, since such a project admits that in blind spot cases the actually operative system produces a manifestation that is either free despite an closed modal profile or unfree despite an open modal profile, it must seek the control grounding property in something other than that modal profile. The idea that systems produce different manifestation types on different occasions already suggests where to look: The distinguishing property must be found in the actual etiology of the action. If the same system with the same modal profile can produce both free and unfree actions,

then the difference must be in *how* it produces those actions, and plausibly in how this etiology is involved in explaining them.

With this idea in mind, we can go back and ask ourselves whether the project I am imagining is committed to denying that any form of reasons-responsiveness grounds control. I think that it is not. The point of the preceding discussion was not to question that agents are free in virtue of being responsive to reasons, it was to point out that this sense of 'responsive' cannot be given an RR* specification. What we need to find is the sense of 'responsive' which designates the exercise of a capacity to respond to reasons for the action the agent actually performs.

For a preliminary illustration of this sense I find instructive what Ira M. Schnall writes about the significance of *RR**:

> To say that S, in deciding to do V, lacked WRR [Schnall's version of *RR**, D.H.] is to say that S's decision-making mechanism M was such that even if S were to have had the most rationally and morally compelling reasons not to do V, nevertheless S, using M, would not have decided not to do V. This counterfactual statement tells us about what would (or rather would not) have happened if certain aspects of the case had been different. What does it say about S's actual decision to do V—about how it was made, and in particular, about whether it was made in a rational, irrational, or non-rational manner? Given only the truth of this counterfactual statement, for all we know, all the logical and psychological steps in the process whereby S arrived at the decision to do V, and all the connections, or relations—logical, causal, and otherwise—among all the elements in the various levels of the process, may have been characteristic of the kind of paradigmatically rational decision-making for which agents are morally responsible. (Schnall 2010, p. 279).

Schnall is here proposing a more radical account, an account that does not rely on essentially modal properties at all. What he proposes is a principle like:

> **Control$_4$**: An agent S is in control of their φ-ing because S *exercises* the *local capacity* to respond to reasons for and against this specific type of action.

It is this idea of exercising capacities as central to freedom that I will pursue in the next chapters. The idea is meant to entail no modal dependence claim whatsoever and therefore constitute a more radical actual-sequence approach. However, this ambition might seem impossible to fulfil or at the very least not fulfilled by control$_4$. This is why in the next chapter I need to take some rather unusual steps to show that exercising a

capacity can be spelled out using only features of the actual sequence. The rest of the thesis engages in the project of developing such an account.

Chapter 3:

The Pertinence Problem, Or: How to Demodalise Free Will

1. Introduction

The previous chapter established that traditional accounts of orthonomy still face trouble with cases that follow the spirit of Frankfurt's original insight, the insight that an agent's orthonomy is exclusively grounded in features of the actual sequence. These accounts are still faced with Frankfurt-type cases because they all still adhere to some type of modal dependence claim. They still spell out an agent's orthonomy in terms of alternative possibilities. Consequently, they all face cases in which these alternative possibilities are absent and yet the agent's action is free.

The recalcitrance of Frankfurt intuitions is damaging for those orthonomy accounts that aim to overcome them using modal apparatus. But it is, in addition, methodologically puzzling for accounts that were at least *trying* to adhere to Frankfurt's original insight. Accounts like that of Fischer & Ravizza (1998) are explicitly constructed to capture the intuition that what matters to free agency is the actual sequence. Yet, my cases in the previous chapter undermine Fischer and Ravizza's account all the same. And they undermine it by simply exacerbating the conditions in original Frankfurt-cases. Fischer and Ravizza then seem to have failed to detach their account thoroughly enough from reliance on alternative possibilities. They fail to deliver an *actual* actual sequence approach.

But if even accounts explicitly dedicated to accommodating Frankfurt's insight fail to accommodate it, we might feel at a loss as actual sequence theorists. How do we remove the reference to alternative possibilities from our account of reasons-responsiveness? How do we spell out the sense in which someone *responds* to reasons, that is, without referring to alternative possibilities in which they respond differently?

This chapter takes the first crucial step towards constructing an account in which alternative possibilities are truly absent (and in which the original actual sequence grounding claim is therefore honoured).

The key, I shall suggest, is to recognise that we have been too narrowly focussed on the free will debate alone. The incompatibility between traditional accounts of responding to reasons and the Frankfurtian grounding claim is in fact just the species of a genus of a general conceptual opposition that lies beneath a whole host of philosophical subdisciplines. This opposition is that between *explanationism* and *modalism* about the notion of *non-accidentality*, which is crucial to – at the very least – all orthonomy concepts, like knowledge, perception, and reasons-responsiveness (to name just a few). Non-accidentality denotes, roughly, the orthonomous relationship we must have to the world in order for the relevant notions to apply to us. It expresses, for example, that when we know that p, this was an accomplishment on our part, that when we respond to a reason, this was due to us – and not just merely a coincidence.

Despite the essential role non-accidentality plays in orthonomy notions, it has never quite received the attention it deserves.[49] Instead, a default stance taken towards it across many fields in philosophy is that it is clearly a modal notion to be spelled out in terms of the relevant ranges of alternative possibilities. However, despite this default stance, orthonomy concepts across a wide range of subdisciplines have long faced problem cases which exploit the deep intuition that alternative possibilities are just not pertinent to whether a given connection is non-accidental. These cases have the same structure and follow the same spirit as Frankfurt-cases. They rely on a set of intuitions about non-accidentality that reveal a competitor analysis to the default stance. This analysis holds that non-accidentality is an *explanatory* concept.

The opposition between these *modalist* and *explanationist* approaches is the real reason traditional actual sequence accounts fail. They fail, that is, because while Frankfurt's intuition recommends an explanationist approach to non-accidentality, they spell out non-accidentality according to the modalist tradition. This internal conflict has not been seen so far because it can only be recognized once we zoom out and see the full extent of the problem, which I shall dub the *Pertinence Problem*.[50]

[49] An important precursor to analyses of non-accidentality are general analyses of luck (see Coffman 2007; Pritchard and Whittington 2015), precursors of which are treatments of moral (Nagel 1979; Zimmerman 1987) see and epistemic luck (Pritchard 2005). Luck is a broader phenomenon than accidentality as I use it here, however. Most importantly, luck includes what might be called 'preselectional' cases – cases in which it is an accident that p and q (and not x and y) are connected in the way they are, irrespective of whether their connection itself is robust in a non-accidental way. Non-accidentality is therefore a much narrower phenomenon than luck, because non-accidental connections can still be 'preselectionally' lucky in a variety of ways.

[50] Another factor that has obscured the centrality of non-accidentality for the philosophy of orthonomy is a widespread reliance on the phrase 'in the right way' in the literature, which is in effect an anti-accidentality clause the allows the theory in which it is embedded to just ignore all problem cases that have to do with accidental connections. To name just two relevant examples: 1. Fischer and Ravizza (1998) rely on the notion in their account of reasons-responsiveness, adding in a somewhat guilty tone: "Of course, it is a notorious problem in action theory to specify what this 'appropriate relationship is' [...] we do not have a specific proposal here". (Fischer & Ravizza, 1998, 64, fn 4.).

Recognition of the Pertinence Problem opens up a new avenue for actual sequence theorists. The key to exorcising the problematic alternative possibilities from the notion of reasons-responsiveness is to reject the modalist conception of non-accidentality. An *actual* actual sequence approach can only be developed by providing an explanationist account of non-accidentality. In fact, the provision of just such an account will be what occupies much of the remaining thesis.

Trying to clear dialectical (or indeed logical) space for a new position is tricky though, especially in an underdiscussed area in which the default position has a firm grip in conceptualising both intuitions about cases and the dialectical situation. My case for recognising explanatinionism as a proper competitor position about non-accidentality will therefore be somewhat bold. The chapter is designed to show that if we accept the modalism vs. explanationism distinction, we will gain a powerful conceptual tool for understanding why the same range of cases and oppositions keep coming up in different subdisciplines. My case for the distinction is therefore to grant it as a working hypothesis and confirm it via its explanatory potential. In this chapter, the case is only made for explanationism being a viable *option*. I will be able to prove the full explanatory potential of the view only with a concrete explanationist account in hand, which I will develop in chapter 6. My reader is welcome to jump to that chapter for details and continuation of some thoughts that come up here already.

Here is the plan. While section 2 provides some preparatory work, sections 3 and 4 introduce a general pattern, which I think reasons-responsiveness and many other notions exhibit. To solidify this hypothesis, in section 5 I show how the pattern comes up in two example debates: that surrounding *knowledge* and that surrounding *moral worth*. Section 6 reapplies the insights from the previous sections to the free will debate and section 7 charts a path for the novel explanationist account of reasons-responsiveness that I suggest is possible in this chapter.

2. Responding to Reasons and Non-Accidentality

One result of the previous chapter was that for an action to be free, it is not enough that the agent merely *possesses* the capacity to respond to reasons. They must also *exercise* this capacity. They must actually respond.

2. Carolina Sartorio (Sartorio 2016) develops a very promising actual sequence approach in terms of absence causation, but relies on the phrase when distinguishing behaviour merely caused by (absences of) reasons from *responses* to reasons, adding defensively: "The distinction between deviant and normal causal chains is a common distinction in the philosophy of action literature." (Sartorio 2016, 135).

But what is a *response*?

In answering this question, it is helpful to look more broadly at the *type* of notion responding to reasons is. Orthonomy is about the ability to get things right and get them right normatively. But reasons-responsiveness isn't the only notion we use to express the range of these abilities. In general, orthonomy notions encompass those capacities that connect world and mind, like knowledge and perception, but also capacities to get things right in more particular normative subdomains, such as the potential for acting in a morally worthy way, i.e. acting based on the salient *moral* considerations in a situation. It might appear to some that these notions share only superficial similarities. But, as I hope this chapter and this thesis will show, in fact the opposite is true. These notions share essential similarities that allow us to group them under the heading 'orthonomy notion' and treat them in structurally similar ways. This is because, at base, they all express a special type of agential success – they encode accomplishments or achievements on the part of the agent. Accomplishment in getting things right about a domain, in turn, expresses that the agent stands in special relationship with the world. What unites orthonomy notions is this relationship – and the puzzles and problems it brings.

Hence, the concept of a response, which is here being used as a semi-technical term to latch onto the relevant phenomenon, expresses a relationship between reason and action/attitude. To respond to a reason is to be in contact with it in a way that has impact on the action or attitude we count as a response. This is after all the core idea of orthonomy: that freedom comes from exercising the ability to adjust our actions and attitudes to the normative landscape. Crucially, it is not enough for this adjustment that our actions and attitudes just happen to correspond to the relevant reasons. This would be, as it were, a mere living in parallel with normative reality. If I run downstairs for no particular reason and it turns out that the hotel was on fire, then there was a strong reason to do what I did, but I did not do it as a *response* to this reason. What the notion of a response expresses is then not just that our actions and attitudes corresponded or conformed to reasons, but that they did so non-accidentally. To say that I responded to a reason in action is to say that it wasn't merely a coincidence that I did what the reason recommended. To respond is to do what the reason recommended non-accidentally. At the core of the notion of a response is then what I shall call a *non-accidental connection* between reason and action.

The story I have just told about the notion of a response can be told – and has been told – about many other crucial orthonomy notions, notions that are about our abilities to get things right. In this chapter, I shall mainly focus on two especially obvious orthonomy notions to point out structural similarities: knowledge and moral worth.

What is knowledge? Clearly, the concept expresses a relationship between an agent's beliefs and the truths they are about. And again, to know that p is not just to believe that p and p being true. For agents might believe that p and lack the kind of relationship to truth that knowledge requires. If I believe the hotel is on fire because a brick hit my head, and the hotel is in fact on fire, then I believed what was in fact the truth, but my belief does not constitute knowledge. What the notion of knowledge expresses is then not just that my beliefs corresponded or conformed to reality, but that they did so non-accidentally. When I have knowledge of p, then it isn't just a coincidence that I believe that p in a case in which p is true.

I hope you see that the stories told about reasons-responsiveness and knowledge are the same. We are asking about a special relationship, and what is special about it is that it constitutes a non-accidental connection.

Finally, consider the property called *moral worth* in the literature. This notion expresses a relationship between an agent's motivations and the properties that make their actions morally right. Once again, for an action to be morally worthy, its motivation cannot just happen to align with what is morally required. If I jump in to save the drowning child but my motivation was to impress my crush Wendy, then I will have done the right thing accidentally. When my action is morally worthy, then it isn't just a coincidence that I was to motivated to do it and it was right.

I will investigate these cases and notions in much more detail throughout the next chapters as I offer an explanationist account of non-accidentality. Here, all I want to point out is that the three narratives are structurally identical. They are about a particular kind of connection to the world, and the key concept that guides intuitions about this type of connection is that of non-accidentality. Moreover, as their underlying structure becomes visible, we can find similar cases and narratives all over philosophy.[51]

The recognition that all these notions require the same kind of relationship is significant. For if the same kind of relationship underlies orthonomy notions, then it will not be surprising if the analyses of these notions share the same woes. If there is a general problem with the analysis of the underlying non-accidental relationship, then we should expect this problem to haunt the corresponding orthonomy notions too. The next section aims to show that there is indeed such a general problem with non-accidentality.

[51] This includes the candidates I have already namedropped: action, responding to reasons, following rules, moral worth, knowledge, perception. But interestingly, it also includes some surprising candidate notions. Ned Hall's (2004) seminal paper *Two Concepts of Causation* makes an almost identical point about causation (although his conclusion is that the notion is ambiguous between the modalist and the explanationist reading). Lange (2016) points out that concepts in mathematics, such as that of a proof, exhibit sensitivity to cases of accidentality.

3. Modalism and Explanationism

My summary of the core intuition behind what it means for an action/attitude to be a response was almost entirely non-committal about what type of relationship is required. All I said was that it must be a relationship to 'the world'. The distinction that will prove crucial to the view developed in this thesis is about what type of relationship non-accidental connections are.

Before we get to the distinction itself however, I need to indicate another point of agreement across subdisciplines about what type of relationship it is *not*. Talking about orthonomy as a way of relating to the normative and non-normative world suggests that non-accidental connections must be some type of causal connection (how else would we be able to get things right?). However, a well-known class of cases which all orthonomy notions face shows that while causal connections might be necessary for non-accidental relationships, they are certainly not sufficient. The examples I have in mind are cases of causal deviance, which I already briefly discussed in the previous chapter.[52] There, they came up as supporting exercise views of reasons-responsiveness. Here, I will discuss their more general lesson. The phenomenon is essential to my eventual account however, so it is going to be a recurring theme across the thesis.

To see the impact the cases have in the current context, consider the following examples:

> **Climber**. Miriam knows her climbing buddy is secretly plotting to drop her into a crevasse and right about now is the last time frame to foil their plans. Coming to realise this fact causes Miriam to become incredibly nervous, which causes her to drop her climbing buddy.[53]
>
> **Fire**. The smoke from his burning hotel room causes Felix to hit his head, which causes him to believe that his hotel is on fire.

In these cases, agents are in causal contact with the considerations that constitute reasons for them – the plotting buddy and closing window of time, the smoke as an indicator of fire –, but the agents' action/attitudes are not responses to these reasons. Miriam doesn't *respond* to her reason for dropping her buddy when she drops them. Felix is not responding to his reason for believing that the hotel is on fire in his belief.

Importantly moreover, both reactions aren't responses because they still seem accidental. It still seems accidental, that is, that Miriam does what her reasons

[52] For treatments of the problem of deviance, see Peacocke (1979a/b), Brand (1984), Bishop (1989), Mele (2003), Schlosser (2007, 2011), Wu (2016). For the best currently available discussion of the problem of deviance, see Mayr (2011), ch.5. I discuss the problem more extensively in ch. 6, section 5.

[53] This is obviously a variation of the famous case presented in Davidson (1963).

recommend, and that Felix believes what he has reason to believe. The problem of deviance is an accidentality problem, at base.

Causal contact to reality is therefore not enough for a non-accidental relationship. At the very least[54], something more is required. Non-accidental relationships are special. What we need is an account of their specialness.

I want to suggest that there are two ways of thinking of non-accidentality at this stage. Since non-accidentality is often only treated implicitly through some orthonomy notion, these approaches are for the most part not explicitly held. But my case studies below shall provide evidence for their background influence. In order to describe these approaches, allow me to introduce the following way of expressing things: I shall continue to speak of accidental *connections* so as not to presuppose anything, and I shall refer schematically to the items of these connections – actions and reasons, beliefs and truths etc. – as p and q, which can be treated as facts. However, this is merely a grammatical convenience. The connection between the fact that S φ-s and the fact that S has a reason to φ is not supposed to be anything over and above the connection between reason and action. You will also notice that talk of 'connections' can be interpreted symmetrically whereas the non-accidental relationships I am talking about seem asymmetrical. I will address these worries (and many more details) in chapter 6. In the current chapter, I just want to develop the very basic idea that there are two distinct ways to think about non-accidentality.

With this out of the way, we can distinguish between *modalism* and *explanationism* about non-accidentality.

The crucial thought behind modalism[55] is that the mark of an accidental connection between p and q is that p fails to be modally sensitive to q in the right way. That is, the actual connection between p and q is accidental, intuitively, if p would have obtained regardless of whether q obtained. Conversely, in non-accidental connections facts are modal difference-makers, as it were. Non-accidentality must then consist in a kind of modal dependence between facts. For p and q to be non-accidentally connected is for there to be a modal tracking between p and q such that if p had obtained, q would also have obtained.

The idea finds reflection in the cases of deviance discussed above. Miriam's action, arguably, doesn't track her reasons well enough through the relevant modal subspace,

[54] It should be noted that there are also non-causal conceptions of orthonomy concepts, most notably of acting for reasons. See for example Melden (1961), Ginet (1990), O'Connor (2000), Sehon (2005, 2016). These won't feature prominently in this thesis because I find them antecedently unattractive. Sorry.

[55] For modalist treatments of knowledge and moral worth, see the below. For ab almost explicit commitment to modalism about non-accidentality, see Smith (2009), 67-68, Smith (2003).

because her nervousness will make her drop her buddy every time, no matter why she is nervous. Felix would have formed his belief on the basis of hitting his head even if his hotel had not been on fire.

Although I have put these intuitions in a counterfactual guise, a more straightforward and helpful version (as discussed in the last chapter) of a modalist analysis can be presented in categorical terms:

(**Modalist**) The fact that p is non-accidentally connected to the fact that q iff a sufficient proportion of the relevant p-worlds are q-worlds.

You will already recognize (Modalist) as an abstraction of the alternative possibilities accounts of reasons-responsiveness I discussed in the previous chapter. Modalism is attractive because it seems to capture the kind of 'robustness' of non-accidental connections. After all, to say that two facts are non-accidentally connected is to say that they don't just happen to align. They are robustly related to one another. Modalism spells out this idea in terms of modal tracking. Robust connection, the idea goes, are, as it were, local necessities. It is in large part for this reason, that alternative possibilities accounts of reasons-responsiveness are attractive. They latch onto a plausible way of understanding the non-accidentality of responses, according to which responses to reasons are actions that modally track the relevant reasons situation.

I take (Modalist) to be a paradigm representation of modalism. Since I will therefore treat (Modalist) as my modalist deputy in the following discussion, a few comments about it and especially its relation to other possible modalist analyses are in order.

First, (Modalist) is a type of similarity-based analysis. That is, the notion of relevance is to be read as a placeholder which allows the introduction of a similarity measure suitable to the more concrete non-accidental connection in question. By default, the relevant worlds will be those closest to the actual world. Various alternative proposals for the relevant worlds are possible and on the market. But my discussion in the previous chapter went through such proposals for a modalist treatment of reasons-responsiveness, so I will eschew discussion of them here

Second, even though (Modalist) is a categorical modalist analysis, it renders almost the same analysis of non-accidentality as counterfactual analyses would do. This is because under the standard semantics of counterfactuals[56] the sentence "If p had obtained, q would also have obtained" is true at the actual world iff all closest p-worlds are q-worlds. This means that modalist analyses spelled out in terms of counterfactuals that follow the standard analysis will entail (Modalist).

[56] i.e. the analysis that follows the framework set up by Lewis (1973).

Third, and relatedly, let me briefly explain why I choose to focus on the weaker (Modalist): Strictly all-quantified world-sentences are not desirable for a general analysis of non-accidentality. This is because it is unlikely that all or even most instances of non-accidental connections between p and q require a perfect modal tracking between p and q – even if the relevant p-worlds are maximally specified. If I am moderately good at scoring a particular type of goal in table soccer, and in this particular situation I score the goal by exercising my moderate ability to score it, then surely the fact that I kicked the ball in the particular way that I did is non-accidentally connected to the fact that the ball went in the goal. But this connection does not require modal perfection. It does not have to be the case that in all worlds in which the conditions are similar (enough) to the conditions in the actual world and I kick the ball in the way that I do in the actual world, I manage to get the ball into the goal. After all, what it means that I am not a master at scoring this type of goal yet is that sometimes the conditions are right and I try, but nevertheless I fail. So it seems some relevant kicking-worlds will not be scoring-worlds. Yet, the actual connection is non-accidental. The take away from this should be that we only need a *sufficient proportion* of the p-worlds to be q-worlds to secure the kind of modal co-variation that will underpin non-accidental connection according to modalism (see Manley & Wasserman 2008, Jaster 2020).

It might nevertheless be objected that I cannot infer a problem for modalism in general from a problem for (Modalist). After all, other modalist treatments of counterfactuals are available (see for example the hypothetical treatment of counterfactuals in Barnett 2009), as are more fine-grained ways of picking out worlds, as are treatments of non-accidentality embedded in more progressive approaches to thinking about modality itself (Vetter 2015, Jacobs 2010). Let me be clear: What I refer to as modalism here is the treatment of non-accidentality according to the standard approach to modality. The standard approach to modality, I take it, is to spell out modal notions in terms of alternative possibilities. Apostate views on the nature of modality will have a more complicated dialectical relationship with the points that this chapter (and the thesis) develops (some of these views are possibly congenial to the explanationist actual sequence picture I am aiming for).

Modalism, thus conceived, is often assumed as the default analysis of non-accidentality. However, there is an alternative picture available. According to this picture, the relevant relationship of non-accidentality is explanatory, and not modal.

The crucial thought behind explanationism[57] is that non-accidentality is characterised by the availability of a special explanation of the relevant action/attitude in terms of

[57] Explanationist approaches are now being developed in epistemology, such as Faraci (2019) and Bogardus & Perrin (forthcoming). Lutz (2020) offers an explanationist approach to justification.

whatever it is a response to. An action/attitude is a response to reasons, on this view, if we can *explain* the action/attitude in terms of these reasons in a particular way.

Cases of deviance provide a good illustration for what makes the explanationist position interesting and prima facie distinct from the modalist idea. The explanationist will point out that what is missing in Miriam's and Felix's case isn't necessarily a modal connection, but the availability of specific type of explanation of their action/attitude. We cannot explain Miriam's action in terms of a reasons-explanation. Nor can we explain why Felix believes the hotel is on fire in terms of his reasons. To be clear, it's not that there is *no* explanation available of Miriam's action and Felix's belief. It seems part of the intelligibility of these cases lies in the fact that we can tell a causal origin story of their action/attitude. But such *'thin'* explanations aren't at issue in cases of causal deviance. Clearly in these cases a more substantive explanation in terms of the agent's reasons is not forthcoming. The explanationist holds that non-accidentality can be understood through whatever it is that provides us with these more substantial explanations in the relevant non-deviant cases. I shall use the placeholder term *unique explanation* for this for now.

(**Explanationist**) The fact that p is non-accidentally connected to the fact that q iff q explains$_{unique}$ p.

You might be worried that (Explanationist) is only prima facie a competitor to (Modalist), and that the two positions ultimately converge. These worries are the birthing pains of a new position, which I can here only counter by asking you to bear with me. It is true that modal concepts are often used to illuminate various types of explanation, and it has even been proposed that we can give a completely counterfactual or modal analysis of the concept of explanation itself (for example in Woodward 2003, Hillel-Ruben 1994, Reutlinger 2016). The explanationist point is that unique explanation does not submit to such attempts. But since I can only fully unfold this point with a developed explanationist account at my disposal in chapter 6, I will here rely on showing that explanatory and modal intuitions come apart for orthonomy notions across many subdisciplines. This, I hope, will make enough of a prima facie case for the distinction.

4. The Pertinence Problem

The most important way to see that (Explanationist) and (Modalist) come apart is that modalist conceptions have been dealing for a long time with tricky counterexamples. These counterexamples track intuitions about explanation as opposed to modal profiles. They always share the same structure, and can be divided in two types:

> **Type 1**: The fact that p is intuitively non-accidentally connected to the fact that q but it is not true that a sufficient proportion of the relevant p-worlds are q-worlds.

Type 1 examples will typically involve an "evil demon", who, while remaining uninvolved in the actual situation, will intervene in the closest p-worlds, thereby erasing instances of the fact that q and making it false that a sufficient proportion of p-worlds are q-worlds.

> **Type 2**: The fact that p is intuitively *not* non-accidentally connected to the fact that q but it is nevertheless true that a sufficient proportion of p-worlds are q-worlds.

Type 2 examples will typically involve a "guardian angel", who, while remaining uninvolved in the actual situation, will intervene in the closest p-worlds, making it the case that q and thereby making it true that a sufficient proportion of the p-worlds are q-worlds.

I say that these counterexamples to (Modalist) 'typically' involve such philosophically unsavoury characters as angels and demons because, while these characters allow us to cut through the dialectical weeds most easily, they are in no way necessary to create these types of counterexample. What is important is the functional role that these characters play. They play the role of what I will call a *modal exploit* – a device that allows us to control for certain features of modal space while leaving all relevant truth-values in the actual world unchanged. In type 1 examples, the exploit is a *mask* because it masks the important truths in the relevant modal subspace. In type 2 examples, the exploit is a *mimic*, because it mimics the important truths in the relevant modal subspace.

Frankfurt-cases or more generally finking and masking cases in the literature on abilities and dispositions belong to the class of modal exploits, but so do many other devices less obviously recognisable as such. I discuss structurally identical cases below.

There are of course strategies in the literature to accommodate modal exploit cases into the modalist picture. But we have seen how resilient the corresponding pertinence intuitions are to such accommodation already in the previous chapter in the example of alternative possibilities views of reasons-responsiveness. In the blind spot cases of the last chapter, the agent's action and reason are not non-accidentally connected, yet we have already abstracted away from interferences and focussed on intrinsic properties. That is, we have already restricted the relevant modal subspace heavily, and yet we still don't get a non-accidental connection in the actual world.

We can think of the strategies we encountered in chapter 2 as local versions of a more comprehensive dismissive stance towards the idea that non-accidentality in general

might not be a modal phenomenon. My discussion there heavily suggested that there is no way to account for modal exploit examples for the modalist. I will present a fully general argument for this in chapter 6, after the specifics of my preferred account of non-accidentality are clear. But even if there is some way out for the modalist, it isn't in any sense an obvious or uncontroversial solution. So I think we can safely point out that there is a considerable tension between the standard modalist approach to non-accidentality and an important set of intuitions about non-accidentality notions. It is this tension that I will refer to as the pertinence problem. The problem is that while our best – and some might say only – account of non-accidentality is modalist, our intuitions are that modal truths are just not pertinent to the relevant non-accidentality notion. I shall call such intuitions 'non-pertinence intuitions'. So, for any orthonomy notion N, we get two tendencies pulling in different directions.

Pertinence Problem:

(1) N requires non-accidentality.
(2) Non-accidentality is modal.
(3) Modal truths are not pertinent to N.

The problem arises if we are committed to all three propositions. However, if non-pertinence intuitions track explanatory rather than modal matters, the tension can be resolved. For we can then drop (2) in favour of an explanationist account of non-accidentality.

There is clear evidence moreover, that non-pertinence intuitions do track explanatory matters. I expressed this in chapter 1, when I pointed out that in Frankfurt-cases our intuition that the agent is still responding to reasons is guided by the fact that their reason still explains their action. Moreover, what makes Frankfurt-cases interesting is that the counterfactual interveners in them don't seem to impinge upon the availability of such a reasons-explanation. The suspicion is that this feature of Frankfurt cases generalises to all modal exploit cases.

To further substantiate this suspicion and to demonstrate the ubiquity of the pertinence problem, I will look at the two other examples I used to motivate the idea of looking at non-accidentality in the beginning: knowledge and moral worth.

5. Case Studies

5.1. Knowledge

It is possible to have a true justified belief accidentally. This, at any rate, is the received opinion about the central insight that Gettier-cases have afforded us. For instance, say I want to know what time it is and a glance at my trusty kitchen clock tells me it is 10 o'clock.[58] It might indeed be 10 o'clock. But it may further be the case that my kitchen clock has suffered a technical defect lately and is now randomly stuck indicating 10 o' clock. The fact that I believe that it is 10 o'clock then seems accidentally connected to the fact that it is 10 o' clock. Epistemologists have widely agreed that such an accidental connection undermines knowledge.

Unsurprisingly then, we find no shortage of modal accounts of the non-accidentality of knowledge in the literature. I will here focus on the conditions that seems to have garnered the most supporters: the condition that knowledge must be "safe", i.e. a belief counts as knowledge only if it could not have easily been false:

(**Safety**) "A belief that p by S is safe if and only if S would not believe that p on the same basis without it being so that p" (Comesana 2005, 397) (but also Sosa 1999a, Sosa 2002, Pritchard 2012)[59]

The safety condition is motivated by the thought that what deprives my belief of its status as knowledge in the clock-case is that it is false that a sufficient proportion of *belief that p*-worlds are p-worlds. Safety-based epistemology is therefore driven by the conviction that "precarious knowledge", as it were, does not exist. [60]

This is already enough to satisfy the first step of the pertinence problem. Knowledge is a notion that clearly presupposes non-accidentality and the most established way of spelling this non-accidentality out is modal. (Safety) above is formulated counterfactually, but again a slightly more helpful formulation can be given in categorical terms:

[58] This example is originally from Russel (1948).

[59] Others who accept modal conditions of *some sort* include Black (2008), Black and Murphy (2007), Clarke-Doane (2012, 2014, 2015, 2016), DeRose (1995), Dretske (1971), Ichikawa (2011), Luper-Foy (1984), Nozick (1981), Pritchard (2007, 2009), Roush (2005), Sainsbury (1997), Sosa (1999a, 1999b, 2007, 2009), and Williamson (2000).

[60] There is also the contraposition version of this modalist principle, referred to in epistemology as 'sensitivity' (see Nozick 1981, 179), which is now widely regarded as unsuccessful.

(**Modalist - Knowledge**) The fact that S believes that p is non-accidentally connected to the fact that p iff a sufficient proportion of the relevant *belief-that-p*-worlds are p-worlds.

(Modalist – Knowledge) gives us the correct result for my initial example. For since my kitchen clock is stuck indicating 10 o' clock, I would have believed that it is currently 10 o' clock an large fractions of cases in which it isn't in fact 10 o' clock. Large proportions of the worlds in which I believe that it is 10 o' clock, that is, will be worlds in which it isn't 10 o' clock.

However, non-pertinence intuitions are not far behind. For several authors have recently pointed out that cases of non-precarious lack of knowledge and precarious knowledge exist.

First, take a case due to Duncan Pritchard:

> Temp forms his beliefs about the temperature in the room by consulting a thermometer. His beliefs, so formed, are highly reliable, in that any belief he forms on this basis will always be correct. Moreover, he has no reason for thinking that there is anything amiss with his thermometer. But the thermometer is in fact broken and is fluctuating randomly within a given range. Unbeknownst to Temp, there is an agent hidden in the room who is in control of the thermostat whose job it is to ensure that every time Temp consults the thermometer the "reading" on the thermometer corresponds to the temperature in the room.[61] (Pritchard 2012, 260)

Temp does not know the temperature. His true belief seems to be accidental in the relevant way. But safety is satisfied. Safety is satisfied because the hidden agent makes it the case that a sufficient proportion of the belief that p-worlds are p-worlds, by matching p with the agent's belief that p in those worlds. The hidden agent, I hope you see, is a mimic. [62]

Masking examples are also available. Comesaña, for example, presents the following case:

> There is a Halloween party at Andy's house, and I am invited. Andy's house is very difficult to find, so he hires Judy to stand at a crossroads and direct people towards the house (Judy's job is to tell people that the party is at the house down

[61] Pritchard uses examples like these two advance an epistemology that uses *both* anti-accidentality clauses *and* ani-luck clauses.

[62] They have the same structural role as Wendy's sainthood in 5.2.

> the left road). Unbeknownst to me, Andy doesn't want Michael to go to the party, so he also tells Judy that if she sees Michael she should tell him the same thing she tells everybody else (that the party is at the house down the left road), but she should immediately phone Andy so that the party can be moved to Adam's house, which is down the right road. I seriously consider disguising myself as Michael, but at the last moment I don't. When I get to the crossroads, I ask Judy where the party is, and she tells me that it is down the left road. (Comesaña 2005, 397)

It seems like I know where the party is. But my belief could have easily been false. That is, a large proportion of the relevant worlds where I ask Judy where the party is, I will be disguised as Michael, prompting her to tell me something false. So a large proportion of my *belief-that-p*-worlds are not p-worlds. Judy's directive and my long-standing robust plan to disguise myself as Michael serve as a mask.

I want to direct your attention now to *why* we think that Temp does not know the temperature even though the connection between his belief and the temperature-facts exhibits the right modal behaviour, and why we think that I know where the party is even though the connection between my belief and the party-facts does not exhibit the right sort of modal behaviour. Our intuitions about the cases are linked to explanatory matters, it seems (see Faraci 2019). My belief that the party is down the left road can be given a rational explanation in terms of the fact that the party is down the left road (and reasons that indicate this truth). Most importantly, the connection between the party-facts and my belief isn't deviant, which is why the relevant explanation is forthcoming. It is perhaps somewhat less obvious how intuitions in Temp's case are also guided by explanatory intuitions. To untangle the case, focus on the fact that Temp's thermometer is broken. Because the thermometer is broken, it will be coincidence that what the thermometer tells Temp is also what is true. So even though the thermometer relays the truth to Temp, it will not relay the truth in a way that makes available an appropriate rational explanation of Temp's belief in terms of the temperature-facts. Temp, in fact, is just like Felix from the deviance case above. There is a causal explanation available of his belief, but a more substantial rational explanation is lacking. This is because the thermometer is merely playing the role of another link in the causal chain that leads to Temp's belief. In a case in which the thermometer *isn't* broken, the explanatory role of the thermometer is not only to causally relay information, but to do in in a specific way. What this way is exactly will be developed in my discussion of the process-specificity of dispositions, and the relationship between deviance and modal exploits (hint: I think masking and mimicking are basically forms of 'modal' deviance) in chapter 6. Here, we can rely on the intuitive judgment that there is some explanatory property that isn't

forthcoming in Temp's case but would be forthcoming if the thermometer wasn't broken. It is this difference in explanation that drives our judgment that Temp does not know the temperature.

The clash between modalist models and explanationist intuitions in these cases will often be expressed in terms of 'relevance' or 'pertinence', which is why I dubbed them non-pertinence intuitions above. For example, in their attack on the safety condition, Hiller and Neta (2007) say about the protagonist of their mimic cases:

> The veritic luckiness of her belief's truth—so understood—is simply irrelevant to whether she has knowledge. (Hiller and Neta 2007, 308).

They address the structure of the pertinence problem as it arises in epistemology pretty directly later:

> Such difficulties seem to attend any attempt to explain what knowledge is in terms of what is true across various possible worlds. Although we accept that ANTI-LUCK is true, it is not clear how an account of its truth can be given in such terms. (Hiller and Neta 2007, 313)

ANTI-LUCK, as you might suspect, refers to the intuition that the connection between belief and truth must be non-accidental for the belief to amount to knowledge. What Hiller and Neta object to, then, is the idea that this intuition could ever be spelled out modally. They have run up against the pertinence problem, which we can now put in terms of knowledge:

(1) Knowledge requires non-accidentality.
(2) Non-accidentality is modal.
(3) Modal truths are not pertinent to knowledge.

I will now show how the same structure reoccurs for another orthonomy notion: moral worth.

5.2. Moral Worth

It is possible to do the right thing accidentally. If I jump into the pond and save the drowning child, but I do it only to impress Wendy, then saving the child is the right thing to do and I'm doing it, but the fact that I'm doing the right thing is coincidental in a problematic way. The evaluative property my action is lacking in such a case has been called 'moral worth' in the literature. Roughly, for an action to be morally worthy is for its motivation to be attuned to the normative situation.

Non-accidentality is crucial to moral worth. That is, in order for my action to have moral worth, it is crucial that the fact that I'm performing it and the fact that it is the right thing to do are non-accidentally connected. These conditions are perfect for the manifestation of the pertinence problem. And indeed, all structural elements of the problem can be found quite easily in the debate surrounding moral worth.

Importantly, the agreement on the non-accidentality condition being crucial has been widespread across opposing accounts of moral worth. According to Kantian views (Sliwa 2016), an action has moral worth iff it is motivated "by the motive of duty" or, in more modern terms, by the fact that it is right where this is to be read as a de dicto motivation. According to Responding views (Arpaly 2002, Markovits 2010), an action has moral worth iff it is motivated by the features that make it right – unsurprisingly called the right-making features. But both Kantian and Responding views agree that the crucial feature that must be captured by a successful account of moral worth is its non-accidentality condition (Baron 1995, p.131; Stratton-Lake 2000, p.56; Herman 1981, p.6; Sliwa 2016, p.6 & p.8; Arpaly 2003, pp. 76–77; Markovits 2010 p.241, Isserow 2018), a convergence helpfully summarised by Johnson King as: "[…]one of the only claims about the nature of moral worth that enjoys anything like widespread consensus across the historical and contemporary literatures on the topic." (Johnson King 2019, p.4).

There is another implicit assumption about non-accidentality widely shared however, namely that it is a modal notion. This is most obvious in Sliwa, who puts it in counterfactual terms:

> The thought is that morally worthy actions are motivated in a way that makes their rightness neither "contingent" nor "precarious"— they are counterfactually robust. (Sliwa 2016, 394)

On the responding side, Arpaly's dispositional condition on moral worth, which she calls "moral concern" is motivated by cases in which the range of circumstances under which an agent would be motivated to perform the right action is vastly different:

> Imagine a person who cares so much for her fellow human beings, or for what she takes rightly to be her moral duty to them, that she would act benevolently even if severe depression came upon her and made it hard for her to pay attention to others. Now imagine benevolence's fairweather friend, who acts benevolently as long as no serious problems cloud her mind but whose benevolent deeds would cease, the way some people drop their exercise programs, if there were a serious crisis in her marriage or her job. Last, imagine the person who acts benevolently on a whim. It is Sunday morning and she is awakened by a call from a charity asking for a donation. Our agent thinks, "Why

> not do something right?" and is moved to do something right so long as her credit card happens to be close enough to the bed.[63] (Arpaly 2003, 87)

Arpaly seems to think that the difference between these types of agents, the difference that determines their degree of praiseworthiness is, in the last instance, modal in nature. It is about what the agent would do under a range of circumstances. For she imagines the crucial differences between agent to arise in virtue of what they would do if those circumstances arose.

To be clear, the idea of both of these accounts is that in order for the agent to do the right thing non-accidentally, their motivations (and thereby their actions) must track rightness across modal space. Thus, these views follow modalist approaches to non-accidentality. For they characterise the right actual connection between the fact that S φ's and the fact φ-ing is right in terms of the range of possible worlds in which both facts obtain. Both views can be given more precise shape introducing a version of (Modalist) for moral worth:

> (**Modalist – Worth**) The fact that S φ's is non-accidentally connected to the fact that φ-ing is right iff a sufficient proportion of the relevant φ-worlds are φ-*ing-is-right*-worlds.[64]

(Modalist – Worth) gives us the correct result for my initial example. If I save the child to impress Wendy, then since my motivation tracks what impresses Wendy, many of the worlds in which I save the child will be worlds where saving it is not the right thing to do, because Wendy will be impressed whether or not saving the child is the right thing to do. Depending on the other moral features of the situation, saving the child will not always be the right thing to do after all. Saving it may cause 1000 others to drown or violate some strong constraint. If you find it hard to imagine cases in which it is morally wrong to save a child from drowning, simply imagine a case in which Wendy is impressed only by keeping a promise and I want to impress her. Surely keeping a promise is not the right thing to do in all situations. Sometimes promises have to be broken so as not to violate stronger requirements. In this case, again, my motivation does not properly track rightness. Hence, my motivation will be modally insensitive to what the right thing to do is, which is exactly what deprives my action of moral worth. So the example in the beginning is well accounted for by (Modalist – Worth).

[63] The keen-eyed reader will spot that this counterexample could be handled by the 'on the same basis' clause which is commonplace in epistemology.

[64] An interesting question I cannot address here is whether moral worth works better with sensitivity or with safety conditions. For some discussion, see Sorensen (2014), Howard (forthcoming).

We can already see the pertinence problem looming: Moral worth is a notion that presupposes non-accidentality and the most obvious and default mode of analysing non-accidentality is modal.

And sure enough, non-pertinence intuitions dominate the debate in this domain as well. For an important charge raised against both Responding and Kantian views is that alternative motivations in possible worlds just don't seem pertinent to whether the actual action has moral worth. Markovits (2010) puts this intuition succinctly:

> The kind of appeal to counterfactuals on which both Arpaly and Stratton-Lake rely can lead us astray. It is often a mistake to ask, when assessing the moral worth of some action, "Would the agent have still done that if. . . ?" If a fanatical dog-lover performs a dangerous rescue operation to save a group of strangers, at great personal risk, should we discount the worthiness of his actions because, had his dog required his heroics at the same time, the doglover would have abandoned the strangers? That he would have done so may be a sign of excessive concern for the dog, rather than of too little concern for the strangers—after all, the dog-lover was willing to risk his own life to save theirs. And given that the dog was not present to deflect our hero's attention from the reasons he had to perform the rescue, it seems ungenerous to withhold praise for so admirable an act simply because the dog might have been present. (Markovits 2010, 210)

But – somewhat curiously – Sliwa expresses the same intuition:

> In general, when deciding whether to give an agent credit for an action—including nonmoral credit—we are interested in the motivations that in fact led the agent to act. Consider a chess player who responded to her opponent's move in the right way, thereby saving her queen: she wanted to win and hence to save her queen and she knew how to do it. If that's what motivated her move, we give her credit for it; it's hardly a fluke that she succeeded in responding correctly to the threat at hand. We do not give her any less credit for it just because, had she suffered from a bout of insomnia before the game, she would have been motivated differently and played badly as a consequence: either because she would have been too tired to care about winning the game or because she would have been to fuzzy headed to think of the correct move. (Sliwa 2016, 399-400)

These intuitions should look familiar. They are paradigms of the non-pertinence intuitions I introduced above. Moreover, if we look closely, we can see that they come with their own inbuilt modal exploit examples. The disposition to play badly after a bout

of insomnia is a kind of mask. Presumably Sliwa thinks her chess player is quite often beset by this affliction, so that a sufficient proportion of the close possible worlds will be insomnia-worlds. Hence, a large proportion of the close possible worlds will not be worlds where she makes the correct move. The worry is that since the chess player's response is still explained in the right way by her reasons for so responding, the actual connection between response and reasons still looks non-accidental. What could have easily been the case seems non-pertinent for the case.[65]

This complaint is also behind Markovits example. I take it that the rescue operation that the dog lover is performing is such that his dog would have very nearly needed rescuing at the same time. So in all close possible worlds, the dog is involved and the dog-lover fails to rescue the people. Yet because in the actual world the dog is absent, the dog-lover acts for the right reasons and thus deserves credit. The disposition to prefer his dog is therefore a mask. [66]

Sliwa and Markovits present masking examples. But it is quite clear that mimic examples are also easily available. To see this, take my initial Wendy example and simply add the following condition: Wendy is a special moral saint, who is impressed only by doing the right thing. My motivation to do what impresses Wendy now tracks the right thing across modal space. So it is true that in a sufficient proportion of worlds where I save the child, saving it is the right thing to do. But clearly, the actual connection between the fact that I save the child and the fact that saving it was the right thing to do is still accidental. My actual action still lacks moral worth. Wendy's sainthood is a mimic: it establishes the relevant modal truths without establishing the actual non-accidental connection. It has the same functional features that the hidden agent in the Temp case in the previous section has.

If you think that this example does not work because it violates another constraint on moral worth, namely that the motivation has to have "moral content", let me indulge the reading of moral content under which this complaint is valid and add to the example that I am motivated to do whatever Wendy, a moral saint, tells me is the right thing to do.[67] This motivation surely has moral content. But Wendy's sainthood is still a mimic.

Thus, the structure of the pertinence problem is present in the moral worth literature[68]:

[65] Again, there are modalist responses available, but I already investigated in what way they are ultimately unsuccessful.

[66] The dog lover also bears similarities to weak spot cases discussed in chapter 2: They are modally closed, as it were, but locally open.

[67] This has moral content because the content of my motivation features the content of moral rightness.

[68] Isserow (2018) explicitly recognises this structure.

(1) Moral Worth requires non-accidentality.
(2) Non-accidentality is modal.
(3) Modal truths are not pertinent to moral worth.

The failure to appreciate that the literature has run up against a more general philosophical problem moreover explains some of the peculiarities of the debate. Let me mention two. First, as you might have noticed, Sliwa seems to directly contradict herself, both advocating for a modal approach to the non-accidentality condition of moral worth and for non-pertinence intuitions. Sliwa seeks to amend this apparent contradiction by introducing the epistemic condition on moral worth that the agent must know that φ-ing is the right thing to do. She thinks that the counterfactuals linked to knowledge – the agent could not have easily been wrong about φ-ing being the right thing –, are somehow of a different kind than the counterfactuals she rejects. But it remains unclear in her paper why this would be so. Does Sliwa want to say that the chess player in her example lacks knowledge about what the correct move is in her actual situation? It seems like the chess player could easily have been wrong about what the correct move is, and yet her actual move is not just a fluke.

The problem here is that Sliwa has tried to solve the pertinence problem without presenting a general solution to it. Her strategy against non-pertinence intuitions is to replace one modal condition, the one that directly links rightness and motivation, with another, the one that links knowledge and rightness. But having a clear perspective on the pertinence problem, we can see that the problem is not *which* modal condition to choose, it is that a modal condition is chosen in the first place.

This failure to see the generality of the problem is even more apparent in the second peculiarity of the debate: Markovits ends her discussion of the various cases in which what she calls "contingent circumstances" (Markovits 2010, 212) influence doing the right thing with explicitly admitting that such influences do not always undermine the moral worth of an action. Quoting this section, Sliwa has the following to say:

> Markovits concludes that since there is no principled line to be drawn, we should give up on counterfactual robustness as a mark of moral worth altogether. I think Markovits is right that views she targets—in particular Arpaly's and Stratton-Lake's accounts—do not succeed in drawing such a principled distinction between those counterfactuals that matter and those that do not. But abandoning the thought that morally worthy actions are non-accidentally right completely strikes me as too high a price to pay. (Sliwa 2016, 400)

But the central error here is that Markovits is *not* giving up on the non-accidentality condition at all. What she is giving up on is the idea that this condition can be spelled

out modally. That is, she is siding with her explanatory intuitions about what actions and motivations are still explained in the right way by moral rightness rather than her intuitions about modal flexibility. But unless we recognize the larger underlying problem, this will certainly look like giving up on non-accidentality altogether, especially within the modalist orthodoxy.

Note that again, what the exchange between Sliwa and Markovits shows is that non-pertinence intuitions track explanatory matters (and can be 'awoken', as it were, by focussing on these matters). Contingent circumstances will not influence our judgments of moral worth as long as they don't impinge upon the availability of the relevant explanation of our (motivation and) action in terms of the right-making properties. This is what happens in the dog-lover case: even though the dog lover could have easily failed to save the strangers, what explains why they did save the strangers on this instance is that they saw and acted upon good reasons to save them. The presence of a disposition to irrationally prefer dogs does not, on first sight, seem to impinge upon the availability of this explanation for this instance of saving strangers. Conversely, think again about the case in which I do the right thing because I do whatever Wendy the moral saint does. It isn't true in this case, it seems, that I act for moral reasons, and so the corresponding explanation of my action in terms of moral reasons is not forthcoming. Just like in deviance cases, there is of course a causal explanation of my action in terms of moral reasons. But what is missing is the more substantial rational explanation of my action, in which I act in light of those moral reasons.

The cases of knowledge and moral worth should have shown that orthonomy notions are beset by the same kind of problem structure, which manifests the clash between two ways to think about non-accidentality. Besides the problem cases discussed here, I have of course presented no argument in favour of one side or the other in this clash. It is open to the modalist to insist that there is a version of (Modalist) that *can* overcome the cases that feed non-pertinence intuitions. However, I have in the previous chapter gone through reasons for why such a strategy is at the very least extremely hard to pull off for the notion of responding to reasons. The points I made there generalise. Basically, there are many ways in which (Modalist) can be restricted by paying more attention to what exactly the relevant worlds are. But whatever restriction is chosen, there will be non-pertinence counterexample cases for the restricted version of (Modalist) as well. The argument offered in chapter 6, section 8 explains why these counterexample cases are so persistent. In a nutshell, they stem from a structural problem with modalism that cannot be removed just by restricting the relevant classes of worlds.

But even absent this argument, the recurring problems with non-pertinence intuitions form the basis for thinking that maybe there is something more fundamentally wrong

with modalism than its precise formulation. For taken together, modal exploit cases form the basis of the following contrastive thought: If we keep the explanatory relation between p and q constant and change the relevant modal truths, our intuition that p and q are non-accidentally connected does not change. But if we keep the modal truths about p and q constant and change the relevant explanatory connection, our intuitions about whether p and q are non-accidentally connected will change (in accordance with whether the explanatory relation is flicked on or off). Thus, maybe the problem with modalism isn't the details but the spirit. And instead of embarking on the project of plugging the holes of modalism, maybe we should develop the explanationist thought further and see where it leads. This is in any case what I shall do. The next sections carry out preparatory work for the project.

6. Towards an Explanationist Account of Reasons-Responsiveness

In the previous chapter, we saw that traditional orthonomy accounts still face trouble with cases that follow the spirit of the original Frankfurtian idea: that agents can be in control in virtue of being orthonomous even in cases in which they could not have responded differently. This idea can be expressed as the claim that an agent's orthonomy is exclusively grounded in features of the actual sequence. What the cases in the previous chapter show is that traditional orthonomy accounts still run up against intuitions that track this idea. This is a deep problem for those accounts that seek to use their modal vocabulary to overcome Frankfurt's original intuition. But it is in addition a uniquely perplexing puzzle for those accounts – most notably that of Fischer and Ravizza – that are explicitly designed to capture Frankfurt's original intuition. How are we going to construct an actual sequence approach in terms of reasons-responsiveness if even the 'arch' actual sequence accounts can't seem to shake appealing to some sort of alternative possibilities?

In this chapter, I suggested that the key to successfully exorcizing any problematic appeal to alternative possibilities from actual sequence approaches lies looking more closely at the larger class of notions reasons-responsiveness (and, in consequence, free agency) are a part of. Orthonomy notions like knowledge, moral worth, and reasons-responsiveness are at base about maintaining a special relationship to the world (normative and non-normative). At the core of this relationship lies the intuition that relevant items – actions and reasons, truths and beliefs etc – must be non-accidentally connected.

There are two irreconcilable approaches to non-accidentality however. By default it is often taken to be a modal notions characterised by a modal tracking of the relevant items

– exactly the kind of tracking appealed to by traditional approaches to reasons-responsiveness. But we can also take non-accidentality to be an explanatory phenomenon characterised by the availability of a special type of explanation. The irreconcilability of these approaches surfaces when the verdicts of the modal approach conflict with the verdicts of the explanatory approach. In these cases, we are left with the distinct impression that alternative possibilities just aren't pertinent to the notion we are talking about. As long as the explanatory matters in he actual sequence are settled, it won't matter what happens in alternative possibilities.

The conflict between modalism and explanationism solves the puzzle about actual sequence approaches. As we saw, Frankfurt-cases are modal exploit cases. They track the intuition that as long a reason explains an agent's action in the appropriate way, the connection between reason and action will be non-accidental irrespective of what happens in alternative possibilities. What they, and their iterations we saw in the previous chapter, really favour then is spelling out the non-accidentality of reasons-responsiveness in an explanationist and not in a modalist way. Traditional accounts of reasons-responsiveness rely on alternative possibilities, we can now see, because they adhere to the default assumption that non-accidentality is a modal notion. The clash between these accounts and Frankfurt-like cases merely reflects the deeper clash between modalism and explanationism then. That is, the structures encountered in the previous chapter reflect the larger problem that haunts orthonomy notions, which I dubbed the Pertinence Problem:

(1) Reasons-Responsiveness requires non-accidentality.
(2) Non-Accidentality is modal.
(3) Modal truths are not pertinent to reasons-responsiveness

Those approaches that seek to overcome Frankfurt's original insight side squarely with (2) here, trying to argue that modal truths are pertinent to reasons-responsiveness and free agency. The previous chapter established that this would seem like an extremely difficult endeavour at best, given the tenacity of non-pertinence intuitions. And in chapter 6 I will present an argument that I believe shows that the modalist approach to non-accidentality *cannot* succeed.

But traditional actual sequence approaches commit a more egregious mistake. They are implicitly contradicting themselves. For in siding with the insights of Frankfurt-cases, they side with (3). However, their way of spelling out reasons-responsiveness still relies on a modalist conception of non-accidentality, expressed in (2). But we cannot hold both (2) and (3). They belong to different approaches to the core notion of non-accidentality – in fact, as a subplot of this thesis is going to allege, they belong to two opposing grand

philosophical pictures of how the mind and the world are connected (see chapter 7, part II).

Hence, the puzzle we started from is solved by focussing on non-accidentality. More than that, the focus on non-accidentality opens up a new option in the debate. For we have now identified exactly where the problematic appeal to alternative possibilities enters the picture. It is through the modalist approach to non-accidentality. Hence, a novel actual sequence approach to free agency as reasons-responsiveness is going to advance through the provision of an *explanationist account of non-accidentality*.

Much of the remainder of this thesis is dedicated to developing an understanding of reasons-responsiveness that conceptualises the underlying non-accidental connection between reason and action in explanatory and not in modal terms. The remaining section in this chapter carries out some important preparatory work for such a project.

7. The Outlines of an Explanationist Account

Orthonomy notions express a special type of relationship between mind and world – a non-accidental connection. A characteristic shared by all orthonomy notions that is closely related to non-accidentality is that they all enable a special post hoc agential criticism. That is, orthonomy notions express the kind of ownership over actions and attitudes that directly enables their creditworthiness. I call notions of creditworthiness 'post hoc agential evaluative' because they apply to actions or attitudes only if these actions/attitudes not only satisfy some corresponding deontic standard, but also satisfy it in a sense that makes the satisfaction/violation of the standard attributable to the agent. Recall: When an agent does the right thing accidentally, they do the right thing. So, a deontic standard applies to their actions, yet an essential agential standard fails to apply. When an agent believes the true and justified thing accidentally, they believe the true and justified thing. So the relevant deontic standard applies to the belief, yet an essential agential standard fails to apply. And so on. Zimmerman calls the evaluative dimension that opens up there *hypological* (Zimmerman 2002, Lord 2018), and I will on occasion refer to it under this description. In general, the idea that orthonomy notions express a non-accidental relationship is inseparable from the idea that they encode a kind of agential success – a success that is due to us as agents. In short, orthonomy notions encode agential *achievement*.

One feature of the achievement encoded in orthonomy notions that I shall explore over the next chapters, but which I want to already mention here is that it points to a unique relationship between the non-accidentally connected items. Knowledge is more than just belief that is also true and justified. A response is more than an action/attitude that

also conforms to a normative reason. That is to say: We cannot just combine two separate conditions in order to obtain non-accidentality notions. We cannot understand them as composites, I shall say (see Lord 2018 for the same vocabulary).

Agential achievement, plausibly, can be accounted for with the assumption that in attaining success the agent exercised their relevant abilities. For hypological notions, the relevant abilities will of course be the abilities to pick up normative significance in the world. Hence, a plausible first step towards understanding what it is to be non-accidentally connected to the world is to understand the respective connections in terms of exercises of capacities. That is, a first plausible assumption to make is that non-accidental success in an action or attitude conforming to some deontic standard occurs when the fact that action/attitude satisfied the standard counts as an exercise of the relevant capacity. For example, when I believe that the thing that is true and justified, this will not be merely an accident if my so believing counts as an exercise of the relevant capacity. When I do the right thing, my action will count as non-accidentally right if it was an exercise of the relevant capacity.

This first hypothesis of course fits with what I established in the previous chapter: that in order to respond to reasons, it does not suffice to possess the capacity to respond to reasons, one must also exercise that (local) capacity. Responding is a hypological notion. It expresses creditworthiness for an action/attitude conforming to a deontic standard, i.e. creditworthiness for doing or believing (or any other reasons-responsive attitude) what the reason recommends. My results can therefore be seamlessly integrated with the hypothesis that we can understand non-accidental relationships through the notion of an exercise. Moreover, they fit with the growing literature across a number of orthonomy-related philosophical subdisciplines that relies on the thought that the problem of causal deviance can be overcome via an appeal to the manifestation of dispositional properties.[69] Consider: When the smoke causes Felix to form the belief that the hotel is on fire, he believes what he has reason to believe and the relevant reason (the smoke-related facts) causes his belief. But the reasons-fact causing his belief still does not constitute an exercise of Felix's capacity to respond to reasons. That exercising capacities helps with the problem of deviance shouldn't surprise us. Cases of deviance are cases of accidentality, so if it is helpful to think about non-accidentality in terms of exercises of capacities, the notion of an exercise should help us out in solving the problem of deviance.

[69] Arpaly 2006, 46-48; Hyman 2015, ch. 5; Lord 2013, ch. 4; Mantel 2018, ch. 2 and 8; Marcus 2012; Mayr 2011, ch. 5; Millar 2019; Setiya 2007,23; Schlosser 2011; Smith 2009; Sosa 2017; Stoecker 2003, 313; Stout 1996, ch.3; Stout 2005, ch.6; Stout 2010; Turri 2011, 390-393; Wedgwood 2006, 664–667. See chapter 7 for a discussion.

Notice however that in pointing out these connections I am relying on an intuitive notion of what it is to exercise a capacity, and on our natural conceptual competence when it comes to distinguishing exercises/manifestations from 'non'-exercises. As I will argue (Chapter 7), such reliance is only helpful to a certain extent. In particular, we can safely rely on our understanding of exercises to explain some of the normative features of reasons-responsiveness. But we will need to dig deeper if we want to understand the core non-accidental relationship between reason and action/attitude. For an explanationist account of non-accidentality, the key to understanding this relationship will be in understanding the unique explanatory relation between reason and action.

These considerations point to what we might call a two-layered approach to the explanationist account of reasons-responsiveness which I shall develop in the remainder of this thesis. In the following two chapters, chapters 4 and 5, I will develop an account of responding to reasons in terms of *exercising* the capacity to respond to reasons. The thought there will be that we can distinguish responsive (to reasons) from unresponsive agency by appealing to exercise-explanations – explanations of φ-ing available in virtue of counting that φ-ing as the exercise of a capacity. Chapter 4 argues for an important feature of the notion of exercising a capacity: capacities can be exercised unsuccessfully. Chapter 5 then applies this result in developing a new exercise-based picture of responsive agency. This is the first layer.

The second layer then concerns the explanatory mechanics that underlie exercise-explanations. That is, chapters 6 and 7 will investigate how exactly exercise-explanations – or more broadly accidentality dispersing explanations – work. The question there is what makes these explanations unique. What is it about them that disperses any sense of accidentality concerning the items they connect? Chapter 6 develops a general account of the explanatory properties needed to understand non-accidental relationships as special explanatory relationships. It gives an account of what makes the explanation special. The idea will be that what makes them special is that they are unified in an important respect. Chapter 7 then applies this account to exercise-explanations and thereby to the account developed in chapter 5.

This is the plan. Let's implement it.

Chapter 4:

Exercising Capacities Badly

1. Introduction

In the last section of the previous chapter, I pointed out that the centrality of non-accidentality for orthonomy notions is linked to the fact that they express agential achievement. And achievement, I hold, is getting things right through the exercise of one's capacities. These tenets might suggest the natural additional assumption that exercising capacities is itself a success notion, i.e. that, as Millar says:

> There is no gap between manifestation and success. The manifestation of an ability is the subject's doing what the ability is an ability to do.[70] (Millar 2009, 227).

This success thesis is important for the wider context of this thesis, because it emphasises a general problem for Exercise Orthonomy Views: If orthonomy consists in achievement, then what happens to those instances of agency that aren't achievements, but which we might still want count as *responsive* agency. If free agency is defined via responding to reasons, how do we classify cases of error, that is, cases in which agents *think* they have and act upon considerations that aren't normative reasons? If responding to reasons is a success notion, then responding to 'false reasons' is not responding at all. Consequently, orthonomy views will have to count all erroneous agents as lacking control. The next chapter will assess the cases alluded to here in more detail.

What I shall challenge in this chapter and the next is the assumption that responding to reasons is a success notion in the sense expressed in the Millar quote above. In *this* chapter, I shall present reasons to doubt the general success thesis as it is expressed by Millar. That is, I will present reasons to think that *in general* exercising capacities is not a success notion in the way the Millar quote suggests. These reasons will lay the

[70] In Millar (2019), this is put slightly more cautiously in places. Millar there holds there that "[…] people exercise the ability to do something only if they do, or are doing, that thing.". This might be strictly true, but it isn't equivalent with the stance quoted in the discussion above. It isn't equivalent because even if I am doing a thing unsuccessfully, it is still that thing I am doing. But if I do it unsuccessfully, I won't be doing what I have the ability to do, namely to do that thing successfully, and so the above quote will entail that I am not exercising my ability in that case. See also Millar (2019), ch.6, footnote 5.

groundwork for my account of how error in the domain of reasons-responsiveness works. To anticipate: I think we pick up on and often act upon considerations that are false. These instances will still count as exercises of the capacity to respond to reasons because exercising in general allows for defective exercises. What this chapter seeks to show is that the assumption that there cannot be defective exercises (i.e. that the success thesis is true) will lead to a warped and implausible conception of defective agency.

2. A Note about Factivity

The issues I am about to engage with are sometimes discussed under the label of 'factivity'. We all know the thesis that knowledge is factive. It says that S knows that p only if p - or in other words: We cannot *know* falsehoods.

It is easy to construct the parallel thesis for *responding to reasons*. If I respond to a reason for φ-ing, it can be said, then surely it follows that there was something I responded to, namely a reason for φ-ing. Thus, an agent responds to a reason for φ-ing only if there is a reason for φ-ing. This would be factualism about responding to reasons.

But to rely on this notion of factivity would be to misunderstand the general theoretical project of providing an account of reasons-responsiveness. Factivity is a notion borrowed from linguistics. It is about whether certain natural language expressions carry the entailment of truth. It asks, for example about the expression "it dawned on me that Hildegard hated me" whether it entails that Hildegard did indeed hate me. But the term "reasons-responsiveness" is not a natural language expression. While the phenomenon it describes is a very real and important part of our everyday evaluative practices of attributing creditworthiness, the term we have elected to do the theoretical work of capturing this phenomenon is a technical term. We have to be careful not to confuse the natural language implications of the technical term with what is really important, namely whether the phenomenon itself ought to be described as a success phenomenon, i.e. whether responding to reasons is a phenomenon that does not occur unless it involves the required success. Otherwise, we will confuse features of the model with features of what it is a model *for*, projecting features of the theoretical apparatus onto the real phenomenon we meant to capture.

It is for this reason that in this chapter and the next, I will not engage with the part of the debate about factivity and success that is purely concerned with the linguistic aspects of some of the notions crucial for spelling out responding to reasons. There is a long-standing controversy about the implications of terms like "acting for reasons" and

"having" or "possessing reasons".[71] This debate is about whether there is a linguistic story defenders of non-factivity can tell that, for example, satisfyingly explains why it would be highly marked to say that "John went to room 1.03 because the lecture took place there, but it didn't take place there." Defenders of non-factivity will usually appeal to some form of pragmatic explanation, for example the explanation that we use sentences like the one about John to pragmatically implicate or presuppose that there was such a reason for which John acted (see for example Comesana & MacGrath 2014, sect. 3.2 , Fantl 2014). While this strategy needs further work, it is far from clear which side of the debate comes out on top, nor indeed whether any side can. This is because linguistic intuitions in these cases will be considerably influenced by our background beliefs about what sort of phenomenon we are describing when we are describing reasons-responsiveness.

This is why I think the debate about the linguistic aspects is to some degree a red herring. What we want to understand is the type of agency involved when agents base their actions on reasons. The technical vocabulary philosophy has chosen to describe this phenomenon will not be an unbiased guide towards its metaphysical features. Rather, the metaphysical features will have to determine, in large part, how we regiment the technical language we use to describe it.

The question whether reasons-responsiveness is a success phenomenon is then at least controversial. It should be settled by looking at the features of the underlying phenomenon – a capacity to track and adapt to normative significance in the world -, not merely by looking at the linguistic features of the terms we have chosen to describe this capacity. Adapting this methodology, I will try to show in this chapter that there are very good reasons, based in the phenomenon itself, to favour a non-success conception of responding to reasons. This conception will have some unintuitive natural language implications, which I am happy to live with given that it captures the underlying phenomenon well.

The reasons I have in mind are general and they have to do with how we should think about human error in the context of agency and ability. More specifically, I am here thinking about *exercising* reasons-responsiveness. Hence, we can learn a lot about exercising reasons-responsiveness by thinking about exercising capacities more generally. There are very good reasons to think that exercising a capacity is not a success notion, i.e. that an agent may exercise capacity without thereby achieving the relevant

[71] For major contributions in the debate, see Alvarez (2018); Comesana & McGrath (2014); Dancy (2000); Dancy (2004b); Fantl & McGrath (2009); Fantl (2014); Hornsby (2007); Lord (2007); Littlejohn (2012), Mele (2007); Millar (2004); Miller (2007); Schroeder (2008). The best discussion of these issues is found in a Pea Soup post by Clayton Littlejohn: https://peasoup.typepad.com/peasoup/2010/02/thoughts-as-motivating-reasons.

success determined by the description of the capacity. Why is exercising not a success notion? Because if it was, we could not intelligibly integrate the phenomenon of error in our theory of agency. If this is true in general, it is also true for exercising reasons-responsiveness. So, if it is true in general that there can be defective exercises of capacities, it is true for reasons-responsiveness too. This, at any rate, will be my argument.

3. The Success Thesis

As we saw from the Millar quote above, there is some allure to the thought that exercising capacities is a success notion, such that there are no defective[72] exercises.[73]

I call this thesis (Success):

(**Success**) Any given x-ing is an exercise of the capacity to φ only if it is a (successful)[74] φ-ing.

I think that this thesis underlies a lot of the factualist folklore that is currently flooding the philosophical landscape. [75] It is rarely explicit[76], as still little explicit literature on exercising capacities exists. But I have encountered the thesis at conferences, in talks, and in private conversations; and it often does important background work in accounts that rely on exercise-explanations in their theoretical apparatus.

For example, here is Errol Lord, with whom I'll be engaging a lot in the next chapter, talking about cases in which someone knows when the colloquium starts (therefore successfully tracking a reason to go to Robertson Hall, where the colloquium takes place)

---

[72] There is a purely qualitative reading of defective that isn't at issue, of course, according to which performances can be defective relative to a non-constitutive standard of success. I might not dance very well, for example, yet still dance. This chapter defends the thesis that there are exercises of the ability to φ that are defective in the sense of not even being successful φ-ings.

[73] There is, of course, also a lot of support for the thesis that there are defective exercises. McDowell says: "If a capacity is fallible, or if…anyone who has it is fallible in respect of it, that means that there can be exercises of the capacity in which its possessor does not do what the capacity is specified as a capacity to do." (McDowell 2011, 37)
Virtue epistemology is another place where the Success Thesis is rejected, especially in the work of Ernest Sosa and John Greco (see Greco 2010 and Sosa 2007, 2010)

[74] The bracketed second instance is only needed if we have a conception of agency according to which some action token can be part of the type φ without being the manifestation of a capacity, which is not a very attractive, but, I take it, a possible position.

[75] Miracchi (2015) holds a version of the view. However, she distinguishes between manifestations and exercises. The latter can have defective instances on her view, it seems. Nevertheless, the knowledge-first conception she develops is heavily influenced by the success thesis. Littlejohn (2014b) seems to hold the Success Thesis only for certain types of abilities and explicitly distances himself from Millar (2009), but his factualist ambitions are in part supported by something like the Success Thesis.

[76] At the time of writing the first draft of this chapter, neither Millar's book nor the insightful counterargument in Carter (2019) were published.

and cases where someone is misinformed about the time of the colloquium (therefore, in a sense, failing to respond to reasons):

> On the natural interpretation of Normal Colloquium (The success case), I act for the normative reason provided by the fact that the colloquium starts at 4 pm. (…) I do that when my action is a manifestation of a disposition sensitive to the property of being a normative reason to go to Robertson Hall. *I cannot be manifesting such a disposition in Unusual Colloquium (*the failure case*). This is because there is no fact that colloquium starts at 4 pm that has the property of being a normative reason to go to Robertson Hall*. So I cannot be manifesting an essentially normative disposition with that as its manifestation condition. (Lord 2018, 151, my italics, my brackets)

Here, we see (Success) taking effect. Or rather, we don't see it, but it does take effect. For it occupies the enthymemical space between the first and second sentence of the italicised bit in the quote. Lord jumps from the lack of a true proposition for the agent to the assumption that no "essentially normative" disposition is manifested. But this doesn't follow. It follows only if manifestations or exercises are success notions, that is, it follows only if (Success) is true. As we will see later (chapter 5, section 4), the argument that Lord offers across 4 different chapters of his book stands and falls with the truth of (Success).

I think that (Success) is false. I think that agents can exercise their capacities badly. If they do, they will fail to achieve the external success attached to the relevant capacity they are nevertheless exercising.[77] I also think that (Success) is false in another way. It also precludes the possibility of capacities exercised *well* while exercised unsuccessfully. Both possibilities – subpar manifestations without success and good manifestation without success are important and I will address them in turn. That is, I will defend:

(i) It is possible to exercise the capacity to φ without φ-ing successfully.

(ii) It is possible to exercise the capacity to φ *well* without φ-ing successfully.

[77] It might be thought that my arguments can be pre-emptively disarmed by pointing out surely there is some sense in which some activities can be engaged in without their final causal result coming about. Millar (2019), 135, calls such activities 'completable'. Making an omelette is such an activity. When I go through the steps of making on omelette, I am certainly exercising the ability to go through the steps that eventually produce an omelette – thus, I am exercising my ability to make an omelette. Millar points out that that I can exercise this ability imperfectly in the sense of exercising it incompletely – i.e. I ca just stop before the omelette is successfully produced. But according to him, this does not violate the Success Thesis, because in exercising the ability to make an omelette incompletely, the agent is still doing what the ability is an ability to do. I am not so sure about this reasoning. An ability to bring to completion the making of an omelette is just that: the ability to complete the process. Surely the agent who fails to complete the process therefore exercises this ability defectively in the sense of not doing what the ability is an ability to do. But this issue is beside the point for my current discussion. For my main counterexamples will not be of completable performances. Hitting the target as an Olympic archer isn't like making an omelette. And yet I think an archer can exercise their ability to hit the target in missing that very target.

In general my argument is that (Success) makes impossible a phenomenon we are all familiar with as agents, namely making mistakes *in* action, producing less than perfect manifestations of what we are able to do and being criticisable (often by ourselves) for it. In order to show this, I will first discuss an example and show how the negation of (i) cannot account for what is going on in it. Then I will generalise my argument. After that, I will present a straightforward counterexample to the negation of (ii).

With this, on to the argument.

4. Abilitative Regret

Recall (Success):

(**Success**) Any given x-ing is an exercise of the capacity to φ only if it is a successful φ-ing.

In order to see what is wrong with (Success), consider Rita, an Olympic archer in contest conditions. Rita is a few points away from winning the gold medal, and hitting the bullseye will secure her the medal. However, this is her last shot. Rita flexes her muscles, takes the right posture, draws her bow, fixates the bullseye, and adopts the right breathing pattern. The air is clear, visibility is perfect, there is no wind. Rita then loosens her arrow. But the shot misses. At some point in the pattern, Rita has made some slight mistake. It is often hard, even for the agent themselves, to pinpoint exactly where and how the mistake occurred, and considerable resources will later be employed to find out exactly what about Rita's posture or process of aiming and firing made her miss the shot.[78] But whatever it was, the presumption, especially in Olympic level sports in perfect environmental conditions, will be that it was due to Rita. It was a mistake *she* made in aiming and firing. After all, considerable resources will be employed in order to train her so as to increase the probability of not making this particular mistake again.

Rita will certainly share the impression that the failure to hit was die to her. Anyone who has ever experienced making a mistake in a domain in which they have a high or even moderate level of skill will know the distinct feeling of failing to do what you can do to the best of your abilities. It is a kind of regret fuelled by the realisation that success was within reach in every sense of the term. Nothing was standing in your way, the conditions were right, but you blew it anyway. Let me call this feeling *abilitative regret* for reference purposes. Abilitative regret can be quite powerful and crushing, but it is, it seems to me, at the same time at the very centre of our experience of our own agency. Failures

[78] The lengths to which sports pros and their trainer's go, and the available equipment for analysis, are genuinely fascinating.

brought about by environmental factors are easily discredited as bad luck, so they have little impact on how we experience our abilities as agents. But failures in exercise of what we are good at doing are truly *our* failures, so they constitute instances in which our agency becomes palpable.

We need to be careful not to lose this phenomenon in the philosophical fray. We require a description that does justice to the phenomenon of abilitative regret and its importance for our experience of agency. Here is a description which to me seems to be the correct one, and indeed as I will argue, the only correct one: When Rita aimed and fired, *she exercised her ability to hit the bullseye*. Her shot was a manifestation of this ability. But it was a subpar manifestation. Rita exercised her ability badly. Why do I think that this is a good description? Because it captures what is absolutely essential about Rita's regret: She was already doing what she is good at when she was firing at the bullseye, she was in the middle of exercising her capacity to hit it. But she failed. Rita experienced a failure *in* exercise rather than *of* exercise, as I will call it. Why was it a failure in exercising her capacity *to hit the bullseye* rather than something else? Well, why not? It certainly was what her trainer trained her to do when she was preparing for the contest. It was also what Rita was aiming to do while firing her shot. And it is what Rita is regretful about doing badly afterwards. But I will address this question more extensively below.

Obviously, my description of Rita's situation is incompatible with (Success). But I think this is how it should be because I think Rita's case is one of many that falsify (Success). Proponents of (Success) will want to give a different description of the case. The question then becomes whether they can capture what is important about it without violating (Success). In the next section, I will go through possible alternative descriptions and show that none of them can capture the Rita case. After having gone through some versions of redescribing the case, I will present an argument that shows why no possible redescription could ever work.

5. Alternative Descriptions of Abilitative Regret

Here are some alternative descriptions of the Rita case that I have encountered. Very little literature is published explicitly about theses like (Success), so none of these responses can be found in print (as far as I can see). But they have been put to me at various conferences, events, and in personal conversations.

(i) Rita has the global ability to hit the bullseye but she lacks the local ability to hit the bullseye.

Recall that local abilities, roughly, are abilities agents have in virtue of all or many of the facts about their immediate environment, while global abilities are abilities agents have abstracting away from many or all the facts about their environment (See chapter 2 for more detail). The suggestion then is that Rita does not exercise her ability to hit the bullseye. Instead, her failure to hit the bullseye means that while she has the ability to hit the bullseye in general (globally), she does not have the ability to hit the bullseye now (locally).

But this is a highly implausible suggestion. Surely Rita, a trained Olympic archer in contest conditions *can* hit the bullseye here and now. If she can't, what is she even doing on the field? Moreover, if she wasn't able to hit the bullseye, what is Rita regretful about? If I failed to jump a high fence and it turned out that I was never able to jump it, then I may regret not getting where I wanted to go. But I can't feel abilitative regret, because I haven't failed in doing anything I could have done. I wasn't able to do in the first place, after all. What Rita is distraught about is that she failed to do something she was able to do at a certain time, not that she was not able to do it anyway but would like to benefit from the results of being able to do it and doing it. Her thought is: "Shoot! I could have done it so easily! I was so close!" it isn't "It's a shame I was not able to hit the bullseye"[79]. This also exposes that response (i) is weirdly paternalistic: Rita, the commentators in their booth, and the experts watching the competition surely would agree that Rita was, at the time and place of her shot, able to hit the bullseye. A theory that has the results of attributing systematic error to all of these experts should be treated with suspicion.

So response (i) will not do by a long shot.

(ii) Rita has the local ability to hit the bullseye, but she fails to exercise it.

The second response admits that Rita really *can* hit the bullseye in her current conditions but claims that she does not *exercise* this ability. This response can be supplemented with one of the candidate alternative abilities that Rita does exercise in order to make it more plausible. I will address those candidates under the next heading. Here, I just want to take issue with the idea that Rita fails to exercise her local ability to hit the bullseye.

Admittedly, this response fares better than the first response. But at second glance, it is not a significant improvement. Rita, after all, is not upset because she was doing

[79] Of course she could *say* "I wasn't able to hit the bullseye" to express her thought, but this usage of "wasn't able" is equivalent to "didn't manage to". Managing to do something just means achieving causal success, it seems to me, whether abilitative or not (compare: "John stumbled into the competition, fumbled with the bow and managed to hit the bullseye").

something *other* than hitting the bullseye, even though she ultimately failed to hit the bullseye. She is upset because she failed to do what she was already well on her way to doing: hitting the bullseye.

Rita's abilitative regret isn't about something she *didn't* do, it is about something she did not do well enough, which implies that she did it. In other words, Rita doesn't fail *to* exercise her ability to hit the target, she fails *in* that exercise. But in order to fail *in* the exercise of the ability to hit the bullseye, it must be *this* ability you are exercising.

To see the difference between failure *to* exercise and failure *in* exercise, take another agent in the same conditions, Linus. Linus picks up the bow but suffers a terrible seizure before he can even take position. Linus fails *to* exercise his ability to hit the bullseye. He did not even start to exercise it. He might feel regret about never actually exercising an ability he had (up to the point when the seizure started, at least). But Rita feels regret about doing something badly she was already doing. Rita's failure clearly is different from Linus's failure. Yet their failures are not different if they both fail *to* exercise the ability to hit the bullseye.

But maybe Rita *was* exercising some ability, just not the ability to hit the bullseye. This is what the third response claims.

(iii) Rita fails to exercise her local ability to hit the target, but she does exercise some more general ability, for example her archery ability.[80]

Maybe the implausibility of the second response can be mitigated by pointing out that it is open to the advocate of (Success) to claim that Rita's situation is better described by choosing a more general description of the ability she is exercising. Rita is indeed experiencing a failure in exercise, the thought must go, but the ability she is exercising is her archery ability – a more general kind of disposition to produce good instances of archery. Maybe Rita is feeling regret over not having been the best archer she could have been, instead of feeling regret over not exercising her ability to hit the bullseye successfully.

Again, this response looks promising as long as we don't look at the details of what it is actually saying. One easy way of seeing why there is something wrong with the response

[80] This is Millar's (2019) strategy in ch. 6.4. A basketball player, he argues, may exercise their ability to go through the routine that increases the chance that ball goes in the basket. He concludes: "There is an ability exercised in both successful performances and unsuccessful performances in which the player's routine is implemented. It is the ability to go through the routine, that is, to do what the player generally does to get the ball into the basket—the performance that raises the chance that the ball will go in." (Millar 2019, 135)
But my argument below holds. This ability is not exercised unsuccessfully, even if the basket is missed. Hence, we cannot account, on this strategy, for the distinctive sense of failure we are after (see this section for detail).

is to look at other versions of it. I sometimes encounter the idea that the description of the ability Rita is actually exercising must be a more explicitly "non-success" description. A popular candidate is the ability to *try* to hit the bullseye. Trying to hit the bullseye of course can be exercised perfectly without actually hitting the bullseye. The standards for successfully trying are different from the standards of actually hitting the bullseye, after all. But this is not a virtue of this proposal, it is a vice. For we are here tracking a *failure* on Rita's part. We are tracking the fact that Rita is upset about something which pertains to her own agency. She is upset that she herself has failed in doing something she is good at doing. If we focus on the ability to try to hit the target however, then this crucial element of the case is lost. This is because Rita may make no mistakes whatsoever in *trying* to hit the bullseye and yet not hit the bullseye – that was after all the point of changing the description of her ability in this way. So as far as *this* ability is concerned, Rita's manifestation of it may be a perfect instance of her agency and nothing to be upset or regretful about. Sure, we might say that Rita is upset that the causal results of her trying did not occur. But we will still have missed the point about her situation. For now the situation looks like that of a highly competent Olympic archer in conditions circumstantially highly unfriendly to her hitting the bullseye. If the archer shoots at the target in a hurricane and fails to hit the bullseye, then she will also have failed to hit the bullseye. But unlike Rita, this will not say anything about her agency in the case. It will not be a failure that is in any intelligible sense *due to her*. Instead the failure will be due to the hurricane. But Rita is very much unlike this archer. Her failure is due to her and she knows it, which is central to the phenomenon of abilitative regret. Abilitative regret does not occur in cases where you are merely upset that a certain causal outcome did not occur, but you did all you could have done to the best of your abilities. It occurs when your performance was subpar according to the relevant standards.

Once we see this flaw in the response that works with the ability to try (or similar "non-success" descriptions), we can see that the response that focusses on Rita's archery ability has the same problem, it is just hidden by the vagueness of the description of "archery ability".

What is it to exercise the archery ability? Well, at the very least, the exercise of the archery ability must be such that it can be exercised successfully even though you fail to hit the target. Otherwise, the case will still undermine (Success). The idea must be that in order to successfully exercise the archery ability, Rita only needs to manifest "good archery". We might underpin this by pointing out that the commentators in the booth might say something like: "This was really good archery here today, it is just a shame that it did not suffice for the gold medal".

But once we admit this, the same problem we had with the ability to try arises again. If it is possible for Rita to successfully exercise the archery ability without hitting the bullseye, then where is the failure on Rita's part that is so essential to the situation? After all, Rita is exercising her ability to be a good archer successfully. There is no mistake in this ability, so Rita should be proud of herself. She did everything she could have done perfectly. The causal outcome just did not occur, but this was not due to her agency. So again, we have subtracted away the crucial feature of Rita's situation.

Of course, I have described Rita as making a tiny mistake, something about her posture, her aim or the way she drew the bow must have been slightly off. This might very well be enough to maintain that Rita does not exercise her archery ability perfectly. She is not the best archer she could be. Fair enough, but does this move help the proponent of (Success)? Recall that (Success) says that some x-ing is an exercise of the ability to φ only if it is a successful φ-ing, that is, if it satisfies the standard of success set up by the description of the ability. Given this principle, if Rita is not being a good archer in what she is doing, she is not manifesting her archery ability, because she fails to live up to the standard of success of her archery ability. But the third response claims that Rita *does* manifest her archery ability, so this way of going is not open to the advocate of (iii). In fact, if we gave up (Success), we could say very plausible things about Rita and her archery ability. We might then say that Rita's archery ability has different manifestation types. One of them is hitting the bullseye through her ability. Another one is missing the bullseye but still exhibiting the typical steps associated with the ability (aiming, posture etc.). If this is our picture, I do not substantially disagree with it. But this picture has already conceded my current argumentative goal to me, namely that (Success) is false.

So the third response faces a dilemma. Either it holds onto (Success), but then it will be unable to locate the agentive failure in Rita's situation, consequently failing to explain why Rita feels abilitative regret. Or it describes the archery ability as not conforming to (Success), but this is to concede my point to me: it *is* possible for something to be an exercise of the ability to φ without it being a successful φ-ing.

One last measure against my argument would be to claim that (Success) only holds for some ability types. For example, it is sometimes alleged that some abilities are factive abilities.[81] The possession of these abilities is dependent on there being facts. For example, I have the ability to know that Nicholas Cage is an alien only if Nicholas Cage is in fact an alien (or so people think). Surely, by definition (Success) must hold for this

[81] Spencer (2017), 468 relies on and elaborates the notion of a factive ability, though for slightly different purposes.

class of abilities. The argument would then be that while the archery ability might be non-factive, the ability to hit the bullseye depends on hitting the bullseye.

But this addition to the third response confuses things. I am here not concerned with the conditions under which an agent may possess an ability. I addressed questions related to these conditions in chapter 2. Instead, I want to know what is involved in *exercising* an ability. We have already conceded that Rita *does* have the ability to hit the bullseye, because this is, at the time of her loosing the arrow, the most commonplace assumption we can make. So the discussion cannot be about whether Rita does or does not have the ability to hit the bullseye. The discussion concerns the issue over whether Rita exercises this ability. My claim is that she does, because this explains the specific type of failure that Rita is experiencing. If my opponent wants to claim that Rita does not exercise her ability to hit the bullseye, then they have to provide an alternative explanation. Insisting that the ability to hit the bullseye is factive does not provide such an explanation, it merely insists that (Success) simply *must* be correct. But why? The factive abilities response does not give an answer besides claiming that if Rita doesn't hit the bullseye, she does not exercise the ability to hit the bullseye. This is just a restatement of (Success).

There is of course a larger issue, which surfaces here, that I can only briefly address at this stage. The issue is about how we should think about agents in various stages of 'envatment'[82], their abilities, and in what sense they can be said to be successful. Take a version of Rita who is connected to an upgraded virtual reality machine without her knowledge. The machine interfaces with Rita's nervous system in such a way that it intercepts external stimuli and translates them into the corresponding sensory impressions etc. Moreover, the machine also intercepts and transmits all signals sent from the brain to Rita's body. So what Rita is seeing is a virtual but accurate representation of reality. When Rita picks up and draws her bow, her body is picking up and drawing an actual bow, but all sensory impressions will be simulated, and all movement will be mediated through the machine (i.e. the machine will be its proximate cause).

Let us say she shoots and hits the bullseye. Will she have exercised her ability to hit the bullseye? The advocate of the factive ability proposal will want to say 'no'. At best Rita exercised some internal, non-success-dependent ability. I suspect that the appeal to factive abilities is here tracking the intuition that Rita's success is accidental. It seems accidental that is, that her shot was also something that satisfied a success standard in

[82] The term is meant to reflect the original Putnam (1981), ch.1 example, while making clear that we need not imagine brains in vats to raise the issues the original example raises. The epistemological challenges of examples featuring envatted agents concern the proper relationship that knowing beings must have to the facts known (or the 'external world' as we say). These issues arise even if we imagine, like I do in the present section, agents who are in a significant sense cut off from direct access to the external world.

the real, non-virtual world. All Rita can claim creditworthiness for is that she hit the bullseye virtually. Rita intends to hit the bullseye, that is, and she also hits it. But her intending to hit the bullseye does not have the correct relationship with the world (the worldly success, that is). It is as if Rita is merely living parallel to a world mimicked by her virtual environment, but she is not in the appropriate non-accidental contact with it. And since the exercise of an ability (at the very least of a factive ability) requires a non-accidental relationship to success, it will be false that Rita is exercising her ability to hit the target. This, at any rate, will be the narrative told by the advocate of factive abilities.

I don't think it is obvious that the scenario described is one in which Rita is non-accidentally connected to her success (it is also not obvious that it is an instance of a non-accidental relationship though). One major problem with the case for me is that it is unclear what explanatory role is being played by the VR device. A blind person fitted with the same type of device would, it seems to me, count as perceiving their environment, not as having mere appearances that happen to correspond to reality. And if this person took a shot, their hitting the target too would not count as an accident. In this variant of the case, the device plays an explanatory role that allows us to say that factive abilities were exercised. Perhaps the difference lies in how functionally integrated we describe a mimicking device (this topic comes up in the literature on causal deviance, see for example Bishop 1989, 159 Peacocke 1979, 87; Pears 1975, 66). Or perhaps there is no difference, and the case can be taken to change our judgment about the original envatted Rita. Either way, the original Rita in VR case does not uncontroversially support the thesis at issue, namely that abilities are not exercised at all if they are not exercised successfully (in ch. 6 section 7, I explore this issue more).

So on second glance, the "archery ability" strategy is also unconvincing. It either implicitly concedes my point or it mischaracterises the essential point about Rita's situation. Additions about "factive abilities" also miss the point.

I think that these failures of the various responses I have considered are telling because they are unified. They all face the same type of problem: that of navigating between accounting for Rita's type of failure and (Success). It seems like no matter what course you chart between these two, you will either mischaracterize Rita's failure or violate (Success). But of course this conclusion is stronger than what I've argued for so far. It is stronger because so far, I have only shown that *some* types of response don't go through. There might be others. And I don't have the space or the patience to address them all. Fortunately, I don't need to. I suspect that what unifies the reasons why these responses fail generalises into an argument fit to undermine any kind of (Success) view about exercising abilities. In the next section, I will present this argument.

6. A "Bad Action" Problem for Success Views

What is the problem that a case like Rita's poses for (Success)? In a nutshell, I think the problem is this: (Success) makes the question of whether some action or activity is an exercise of a given ability dependent on whether that action or activity satisfies a standard of success. But it is also a very plausible background assumption that agents' actions are subjects to such a standard of success only if they are exercises of the relevant abilities in the first place. Consequently, (Success) collapses two *separate* (because dependent) standards. It collapses conditions under which a standard *applies* to an action into conditions under which that action satisfies or violates that standard, that is, is either evaluated as good or bad. All views that imply such a collapse have problems with the same kind of case: The kind of case where an action satisfies the first standard – the standard of application – but violates the second, i.e. is a "bad" action in some sense.[83] The reason these cases pose a problem is that if the standards of application and the standards of satisfaction/violation are the same, every action that violates the standard counts as something to which the standard didn't apply in the first place – in which case it makes little sense to say that it violated the standard.[84]

So much for the nutshell. I will now present an argument of this sort specifically aimed at views like (Success), views about what it takes to exercise an ability. I mean the argument to be a kind of reductio. (Success) creates a contradiction when conjoined with two highly plausible theses. To get the spirit of the argument, consider what (Success) says. It says that there is no action or activity that counts as the exercise of the ability to φ and is not successful according the standards of success set up by that ability. But, I object, here are two plausible assumptions: First, there is a standard of success that applies to action or activities only in virtue of them being exercises of the corresponding abilities. Second, it is possible for actions or activities to violate that standard. So it is false that every exercise of the ability to φ is automatically a successful exercise. Here is the argument more precisely:

[83] Bad action problems are usually discussed in connection with constitutivist or constructivist conceptions of normativity such as the Neo-Kantian views presented in Korsgaard (2009). For treatments of the problem in that context, see Clark (2001), Lavin (2004), and Lindemann (2017). Thanks to Olof Leffler for discussions on these issues.

[84] This argument spells trouble for conceptions of the possession of abilities based on the agent's modal success rate like Jaster (2020). I discuss these difficulties in Heering (2020).

(1) There are some x-ings to which a standard[85] of success for φ-ing applies but which violate it. (Assumption)
(2) For all x-ings, x-ing is an exercise of the ability to φ only if it is a successful φ-ing. (Success)
(3) For all x-ings, if x-ing is not a successful φ-ing, it is not an exercise of the ability to φ. (Contraposition from 2)
(4) For all x-ings, a standard of success for φ-ings applies to x-ing only if x-ing is an exercise of the ability to φ. (Assumption)
(5) For all x-ings, a standard of success applies to x-ing only if it is a successful φ-ing. (from 3,4)
(6) There is no x-ing to which a standard of success for φ-ing applies but which violates it. (Contradiction with 1)

My crucial assumptions are (1) and (4). The rest just follows. So let me defend these two premises. Premise (1) claims, basically, that there are "bad" actions. Rita's failure is a bad action. It is a bad shot, a shot which fails to satisfy the standard of correctness set up by the goal of hitting the bullseye. But other bad actions can be found everywhere at every time in abundance. As I have already emphasized, they constitute a crucial aspect of our experience of our own agency. You will not only find the phenomenon manifested in pro golfers missing their put and goal keepers failing to protect their goal, but also in many of our everyday less than perfect instantiations of agency: spilling your coffee while pouring it, making a typo while writing an E-mail, misspeaking, dropping your folder and many, many more. All of these actions fail to live up to a standard of success expressible in terms of the relevant success description of the corresponding ability, for example the ability to make the put, the ability to protect the goal, the ability to pour coffee, the ability to hold your folder.

Let me obviate a misunderstanding of these points. What I am not claiming is that any of the listed actions are intentional under their defective description, i.e. that there are intentional spillings, droppings etc. A spilling is not an action. A pouring is. A spilling is a pouring that is defective, so it is an intentional pouring that does not live up to the pouring standards. So the description 'spilling' does not pick out an action type, it picks out a bad pouring.

So (1) should be familiar to us. The more controversial assumption is made in (4). This premise says that there is a standard of success that only applies to actions if they are exercises of abilities. There might very well be other standards that apply to actions

[85] Technically, this needs to be the identical standard for each premise. But this technicality would complicate the presentation.

merely in virtue of the agent's possession of abilities or in virtue of something other than their agency. I will make no judgment about this question. But it is important that there is also a standard that only applies to actions if they are exercises of abilities. I have argued for this thesis before in Chapter 2. There I argued that possession views of reasons-responsiveness, views according to which it suffices for an agent to possess the ability to respond to reasons for a given action to count as free, cannot be correct. They cannot be correct, recall, because of the phenomenon of deviant causal chains. Even if the agent possesses the most local ability possible, it might be true that the agent possesses the ability to φ and the agent φ-s, but the agent does not φ in virtue of that ability because her φ-ing was brought about deviantly. As I discussed in the last chapter, orthonomy notions like reasons-responsiveness, knowledge, and moral worth, all share this feature because they express agential achievements. In order for an action to count as an achievement, it is not enough that the standards of success for that action apply to it. After all, beliefs can be accidentally true, agents may do the right thing accidentally, and act in accordance with reasons accidentally. Getting things right orthonomously however, requires non-accidentality, which can be expressed by saying that the relevant actions need to be achievements. These features of orthonomy hold for exercising capacities and abilities as well, naturally. If I have the ability to make the putt and I make the putt, but I only make it because I cheated using magnets, then the putt will not count as my achievement and the usual standards of praise will not apply (this is why we don't condone cheating). If I pour my coffee perfectly, this will count as an achievement of mine only if I wasn't guided by an invisible force to pour it perfectly. It must be an exercise of my coffee pouring skills. These conditions for the application of achievement standards may or may not be the only standards our actions and activities are subject to. But they are certainly some of the most important ones. For they are the conditions that make sure our successes and failures are properly connected to our agentive capacities. So (4) is also true. It must be. But (4), (1) and (Success) create a contradiction.

Since (1) and (4) are crucial common-sense assumptions, all the blame for the contradiction falls on (Success). If (Success) is true, then we can never fail to live up to the standards crucial to our agency. We can either be successful in exercising our abilities, in which case achievement is guaranteed. Or we can fail to exercise our abilities, in which case the standards don't even apply to us in the first place, so no real *agentive* failure has taken place. This is an impossibly impoverished, but frighteningly resilient picture of human agency that I hope my arguments here help to overcome.

According to the alternative picture that suggests itself, there are in fact at least two sets of conditions involved in exercising abilities. The first set of conditions determines whether a given x-ing is indeed an exercise of the ability to φ. If an x-ing fails to satisfy

these conditions, it is not an exercise of the ability to φ and so cannot be subject to the standards set up by such exercises. The second set of conditions determines whether a given exercise is a good or bad exercise – with conditions of goodness and badness determined by a variety of factors depending on the ability in question and domain we are in.

This is why we treat competent failures differently from incompetent ones. Competent failures still manifest the underlying ability, they just don't live up to the relevant standards that apply to exercises of that ability. Incompetent failures are often treated with less rather than more criticism because they are judged as too far away from the relevant standards. Those standards don't even apply, the thought often is, so it would be meaningless to use them to assess this kind of failure. We don't judge the guy who just sort of swings the club aimlessly by the standards of golf. He isn't even playing golf yet.[86]

So far, I have argued against (Success) and for the idea that it is possible to exercise the capacity to φ without φ-ing successfully and I have given precise meaning to this possibility. I now want to complicate matters just a little bit more. I will argue via example that it is even possible to exercise the capacity to φ *well* without exercising it successfully.

7. Exercising Capacities Well

How is it possible to exercise a capacity well, but not successfully? This is possible when the standards of external causal success come apart to a significant degree from the standards that make exercises good or bad. Does this ever happen? I think it routinely does.

To see this, imagine a bicycle fitted with a tachometer. The tachometer has the capacity to measure the speed of the bicycle. In fact, if we think of the tachometer as a causal system of the bicycle, then it makes sense to say that the tachometer just is the bicycle's capacity to measure its speed. But for grammatical ease, I shall continue saying that the tachometer has this capacity. The capacity of the tachometer is externally directed. It is the capacity to correctly track facts about the speed of an object and to translate them into a specific scale of measurement.

[86] We are often epistemically uncertain about whether we should treat especially bad exercises as exercises at all. This becomes especially relevant in the case of rationality and politics. How should we deal with people who base their voting decisions on conspiracy theories about Satanist lizard people who harvest the blood of children? Did they exercise their rational abilities very badly or not at all? Our attitudes may change depending on the answer.

Now suppose that I take the bicycle, turn it around, prop it up on a table and spin the wheels. The tachometer will indicate a certain speed. But the bicycle has no speed, so the tachometer is failing according to the external standard of success associated with it. It does not indicate the correct speed of the bicycle.

Yet it seems to me perfectly legitimate to say that the tachometer is exercising its capacity to indicate speed correctly. Moreover, it seems true that it is exercising this capacity well, even perfectly. After all, we frequently use the set-up of this example to test whether the tachometer is exercising its capacity well. We put the tachometer in ideal conditions and see whether its causal system is unimpeded in these conditions. This practice would make little sense if we did not think that this was a way of checking that the tachometer still has and is exercising its capacity to indicate speed. The fact that the tachometer does not actually indicate the correct speed seems entirely beside the point. This is because given the environment the tachometer is in and given its causal make-up as a system, there is no further activity it could perform in order to detect that the environment it is in was set-up to produce factually wrong results. So as far as the tachometer is concerned, it exercises its capacity perfectly. Facts about the environment don't play a role in this judgment. Interestingly, this can also be put in terms of a modal condition: the tachometer is indicating the speed that the bicycle would have if it was in the proper real-world conditions.

There is of course another sense of 'capacity', in which the tachometer lacks the capacity to indicate speed correctly. This is the sense in which we imply opportunity when we speak of ability. The tachometer is in the same situation that Rita the archer would be in if she was linked to a fully VR environment simulating her contest conditions. What I have been arguing for is that even if Rita lacks the opportunity to hit the bullseye, when she is going through all the right steps of exercising her ability to hit the bullseye, she is exercising *this* ability perfectly. We can even imagine future training centres equipped with VR capabilities like those just imagined. If those VR capabilities didn't train the ability to hit the bullseye – and abilities are trained by exercising them presumably – then what other purpose do we imagine them playing in the scenario? The point about the tachometer then is that it is exercising the capacity to indicate speed perfectly – the capacity that it would perhaps lack in interferential circumstances, but which it keeps even in circumstances of lack of opportunity.

I also sometimes come across the response to this example that the tachometer is not in fact exercising its capacity to indicate speed correctly, it only exercises its capacity to count the rotations of the wheels (because presumably this is how the tachometer indicates speed). It seems true that the tachometer is exercising this capacity. But in terms of what other capacities compatible with the exercise of this capacity it might be

exercising, this truth is neither here nor there. Many things have capacities in virtue of having other capacities and exercise these higher-level capacities in virtue of exercising some lower-level capacities. But as long as this relation obtains between them, exercising one capacity is compatible with exercising the other, so the suggestion that if the tachometer doesn't exercise one because it exercises the other is misplaced. If I exercise my capacity to lift my cup, I simultaneously exercise the capacity to flex my muscles in a certain way. But that doesn't mean that I don't exercise the capacity to lift the cup (if anything, it suggests that I *do* exercise that capacity).

The example shows that it is possible and commonplace for exercises of capacities to fail according to an 'external' standard but still succeed according to an 'internal' standard. This complicates matters a little bit, because now there seem to be three different sets of conditions in play: Application conditions on what actions count as exercises, internal success conditions for what counts as a good exercise and external success conditions set up by the description of the relevant ability (what it is an ability *to* do). [87]

This raises the question of how these three sets of conditions are related. We already know that in order for success conditions to even apply to an action (or activity), it must satisfy the first set of conditions (which determine whether some action is an exercise). But the conclusion that it is possible to exercise a capacity well (and therefore be a good manifestation) raises the question whether the inverse is possible, i.e. whether it is possible to exercise a capacity badly but nevertheless successfully according to external standard.

In order to answer this question, I have to say some speculative things about how the internal standards and external standards relate. It seems to me plausible to assume that the two sets of standards are functionally related. That is, if we think of the causal systems that ground capacity-ascriptions as functionally organised, then it is not far-fetched to think that their function is what the relevant ability is an ability *to do*. This is not meant as a reductive analysis of abilities. There is more to having and exercising an ability than merely having and exercising the function of a causal system. But it is plausible to assume that such a function is part of what it is to have and exercise an ability. Our intuitions in the tachometer example latch onto this aspect. The tachometer seems to be exercising its capacity just fine because there is no internal flaw in the causal functional organisation of the system.

[87] This resembles but is not identical to Sosa's AAA analysis of performances in virtue epistemology (Sosa 2007, 22–3). According to this analysis, a performance is accurate when it achieves an external goal, adroit when it manifests a competence, and apt when it is accurate because adroit.

In order to imagine an action (or activity) that does not meet the internal standard but does meet the external standard then, we would have to imagine the causal product of an internally flawed system, for example the tachometer, which has a short circuit, indicating the correct speed. I think what we would usually imagine here is a case where the external success is accidental, for example because the speed indicated just happens to land on the speed the bicycle is currently travelling at. But these cases are not of the right sort. For the most reasonable thing to say about them is that the tachometer does not exercise its capacity to indicate speed in them at all. The correct indication is accidental after all.

Could we instead imagine a tachometer that despite some technical problems, manages in the end to indicate the correct speed (and this is, by stipulation, an exercise of the capacity to indicate speed correctly) or an archer who, despite some trouble with aiming, manages to hit the bullseye? I think if we can imagine such cases, we imagine successful causal outcome that are in some sense *just barely successful*. Maybe the arrow flew in a wobbly line, the needle indicating the speed is bouncing up and down. That is, we don't imagine that an internal functional failure has no impact *whatsoever* on whether the causal outcome is achieved. Moreover, in order to imagine this, it seems we are imagining the functional failure to be almost negligibly minute. It may very well be the case that the chaos of the empirical world makes such scenarios conceivable. But the philosophically idealised situation seems to be that if we violate the internal functional standards, then this will also make us violate the external success standards in situations where both types of standards can be fully and clearly identified and spelled out, which may not ever be the case for our epistemic situation in the actual world.

Just for clarity's sake, here are the three standards:

Application Standards:	Conditions C such that, for a given ability to φ and any action x, x-ing counts as an exercise of the ability to φ iff x-ing meets C.
Internal Standards:	Standards I such that, for any φ-ing (that counts as the exercise of the ability to φ), the ability to φ counts as functioning well iff φ-ing meets I.
External Standards:	Standards E such that, for a given φ-ing (that counts as an exercise of the ability to φ), the φ-ing counts as successful iff it meets E.

The important relation between the three sets of conditions we have found is then this. If a given φ-ing doesn't meet the application standards, then the two other standards

won't apply to that φ-ing. If a given φ-ing violates the internal standard, it also violates the external standard. But a φ-ing may violate the external standard, without violating the internal standard.

Thus, I have argued for my two theses about exercising capacities. It is possible to exercise capacities without exercising them successfully and it is possible to exercise them well without exercising them successfully. So (Success), the thesis that saying that any given x-ing is the exercising of the capacity to φ only if it is a successful φ-ing, is wrong. This now leaves us with the simple task of applying the general results obtained from my discussion to responding to reasons.

8. Responding to Reasons

So far, I have established that there are defective exercises of capacities. That is, agents may exercise their capacity to φ without φ-ing successfully. Responding to reasons, according to my account, is exercising the capacity to respond to reasons. Combining the non-success view of exercising with the exercise view of capacities then renders the result that there may be defective exercises of responsive agency. Responding to reasons unsuccessfully means recognising and acting upon considerations that aren't in fact reasons for the relevant action, either because they are false or because while true, these considerations don't favour the relevant actions.

In other words, my account will sometimes count reactions performed for what the agent thought were reasons as exercises of the capacity to respond to reasons. It will also count some false impressions of agents as to what their reasons are as responses. Those erroneous recognitions of reasons and reactions based on 'false reasons' that the account will count as responsive of course are those that still count as exercising the capacity to respond to reasons, albeit exercising it badly.

Instances of defective exercise of the capacity to respond to reasons will correspond to actions to which the external success standards still apply, but which violate them in some way. There are also instances of exercising RR capacities perfectly, without satisfying the external success standards. In these instances, the internal standards for responsiveness[88] to reasons will be satisfied, but they will significantly come apart from the external standards. You can probably guess that this category is reserved for agents in what is sometimes referred to as New Evil Demon, nor even New New Evil Demon scenarios (Lord 2018).

[88] I will stay silent in this thesis as to what these standards are. I am attracted to the view of Raz (2005) according to which they are basically standards of good reasoning.

Let me take this opportunity to address what might be considered a gap in the argument of the current chapter. The gap is between the assumption that S exercises her ability to φ and the assumption that S is (thereby) φ-ing. We might oppose the notion that we can easily move between these two precisely because of what I've shown so far. An agent hooked up to virtual reality who goes through the typical motions of playing tennis, it might be argued, exercises their ability to play tennis. But they aren't playing tennis. They aren't playing tennis precisely because in order to play tennis, you need a court and a racket – and you need to actually hit the ball. Equally, we might think that exercising the capacity to respond to reasons does not amount to responding to reasons. For the latter requires reasons to respond to (plus perhaps other environmental conditions).

It seems to me that this kind of dialectical impasse cannot be overcome locally, as it were. My view is that the move from exercising the ability to φ to φ-ing is justified because both cases involve the manifestation of the very same capacity. More precisely, as I shall put it in the next chapter, both instances involve the very same explanatory structure, in which we appeal to something being an exercise of the relevant capacity in order to explain it. My opponent's view is that φ-ings differ significantly from 'mere' exercises of the ability to φ as described by me in this chapter. They differ in being linked more substantially to the world. This is why my opponent thinks the move from exercising an ability to φ to φ-ing is not acceptable. Both views are couched in a larger system of thought, from which they cannot be neatly separated, and so cannot be compared in isolation.[89]

In order to garner holistic support for the non-factualist and, as I shall say, *Univocal,* approach that I have started in this chapter, the next chapter will further explore and expand upon the idea that the difference between exercising the capacity to respond to reasons and responding to reasons (as my opponent here would put it) is negligible, i.e. that there is a core phenomenon that can be retained across cases of error and success. Moreover, retaining this core phenomenon – responding to reasons (as I would say) – allows us to explain how a range of cases are systematically connected. It allows us to explain, most importantly, that our evaluative practices will often ignore the difference drawn here between 'exercising a capacity to respond to reasons' and 'responding to reasons'. Responding to reasons is intimately linked to our normative practices of justifying, explaining and blaming in terms of reasons, after all. And normatively speaking, whether in exercising the capacity to respond to reasons the agent actually did latch onto reasons will – perhaps somewhat paradoxically – not make a difference in crucial cases. Perhaps this is something that generalises to all φ-ings, too.

[89] Errol Lord referred to this as a 'spoils to the victor type situation' in conversation.

We don't need a racket and a court to play bad tennis. So whether or not we have a racket and a court won't matter to our evaluation of the playing. The exercise of the relevant capacities is enough.

Chapter 5:

The Exercise Univocal View

Part I
The Positive Proposal

1. Introduction

This chapter develops an account of responding to reasons in terms of the exercise of the capacity to respond to reasons. In doing this, I will rely on and continue to discuss the themes of the last chapter.

The last chapter dealt with a general thesis about how the phenomenon of error fits with the notion of exercising a capacity. But I applied the results of that chapter only abstractly to my target notion of reasons-responsiveness. This chapter engages with the phenomenon of error as it occurs within the domain of responsive agency. I will look at a range of cases of successfully and unsuccessfully responding to reasons. With the help of the results of the last chapter, I will develop a novel account of how these cases are related. This *Exercise Univocal View*, as I shall call it, holds that both in cases of error and success the very same relation between reason and the agent's φ-ing holds, namely that of exercising a capacity.[90] This account allows us to say about various cases what we should like to say. Because the chapter is long, it has two parts. The first part develops the positive account. The second part is spent dissecting why other accounts will invariably mischaracterise either cases of error or of success in some way. The weary reader can take a break after part I.

The main opposition that I will deal with in this chapter is that between Univocal and Non-Univocal Views (Lord 2018, ch.6.4) of responding. Put very roughly, Univocal Views hold that there is one unified phenomenon of responding to reasons and one relation associated with it. This phenomenon is instantiated both in cases of success and in cases

[90] I will speak of exercising capacities and responding to reasons as relations because a) they are best conceived of as relating dispositional properties to φ-ings (for capacities) or reasons to actions (for RR) and b) following the explanationist ideas of chapter 3, I will here focus on the explanatory relations expressed in the concepts of responding and exercising a capacity.

of error. Advocates of Univocal Views usually hold what has been described as a 'factoring' picture (see Cunningham 2019a, 2019b; Lord 2018, 2010; Schroeder 2008 for further discussion), the picture that we can construct the relation in the success case from the relation in the error case by adding a success component.

Non-Univocal Views on the other hand hold that responding is radically split. There are on the one hand cases of successfully responding through the RR capacity, which constitute a kind of achievement.[91] There are on the other hand entirely separate cases of error about reasons, which involve a different relation with a different analysis. Neither of these relations includes the other as a factor. We cannot construct the relation present in cases of success by taking the relation present in cases of error and adding success.

My Exercise Univocal View rejects the factoring picture to some degree, but it also rejects the idea that the phenomenon of responding is radically split. We cannot construct the relation of exercising from cases of non-exercise. But exercising a capacity itself does not entail that there is a sui generis achievement relation in success cases. In both cases of success and error, agents exercise their capacity to respond to reasons.

The discussion will inevitably also touch on what mental phenomena are involved in responding to reasons. I have said already that responding to reasons means both recognising and reacting for reasons. I shall say more, throughout this chapter, about what that means. It should be noted from the outset though that I am still treating exercising a capacity itself as a primitive notion that we can understand well enough for now. So what I will have to say about recognising and reacting for reasons will fall back in crucial places on the notion of an exercise – that is why the view is an exercise view, after all. Questions regarding what it is conceptually and metaphysically to exercise a capacity will have to wait until chapter 7.

Let me now introduce a range of notions and cases that will prove important to this paper.

2. A Spectrum of Cases

To respond to a (sufficient or decisive) reason is to recognise that reason, and to react appropriately for the reason so recognised. These two conditions are not independent. But it is necessary, for the purposes of exposition, to treat them separately. The following cases and distinctions therefore in fact apply equally to the recognitional part of exercising the capacity to respond to reasons, but I will discuss them mainly in terms of

[91] This sense of achievement does not necessarily involve effort, opposed to how Braford (2015) defines the term.

the reactive part. I will also talk about actions mainly, with only one prominent case of belief, which of course, among many other mental states, can count as reactions to reasons.

In order to explore the notion of reacting for a reason, I want to introduce a spectrum of cases, which are going to serve as a way to structure the discussion about the boundaries of the notion of responsiveness. That is, I shall later assess these cases in terms of whether and to what extent they involve responding to normative reasons, and I will discuss how to group them together in a way that makes most sense extensionally.

Let us start with a case of paradigm achievement.

When an agent successfully reacts to a (sufficient or decisive) reason, the agent will act 'in light of' (Dancy 2000, 103; Alvarez 2009b) that reason, that is, they will react *for* this reason. I take it that when an agent reacts in the light of a consideration, this consideration will explain their reaction in terms of a rationalising explanation – an explanation which renders the explanandum minimally rational from the agent's perspective.[92]

When an agent acts for a normative reason, they act based on having recognised the normative significance of a consideration. Consequently, we can explain their action in terms of that normative consideration they recognised. As Lord (2018, 81 ff) points out, we have here what can be called an *achievement explanation*. The agent exercises her capacity to respond to reasons and she gets it right. This constitutes an achievement on her part. She has responded correctly to the normatively significant features of her environment through her own agency. She is liable to be praised and commended for the attitudes and actions which constitute reactions for reasons in this sense. Here is such a case, together with the important truths that hold in it.

> **Achievement**. Rita smells the smoke in the air, can see the flames engulfing her hotel room. She can feel the heat emanating from the walls; the fire alarm is blaring. Rita, who has checked these phenomena carefully for their veracity having found no reason to be suspicious, believes that the hotel is on fire. She leaves the hotel on the basis the fact that the hotel is on fire (which she believes).

[92] This notion is to be treated distinct from any more substantial notion of rationality. There is a extensive debate, which I take to be dialectically downstream from the debate in this chapter, about whether the concept of rationality should be understood in terms of responsiveness to reasons (Kiesewetter 2017; Lord 2018), an idea that used to be commonplace in metanormative theorising (see Smith 2007), or in terms of mental coherence (Broome 2007, 2008. 2013), expressible in (the satisfaction of) so called requirements of rationality. The Exercise Univocal View does have impact on this debate, because it offers a superior analysis of rationality in terms of reasons-responsiveness. That is, it offers the theorem that rationality consists in exercising the capacity to respond to reasons, which, most importantly, solves the problem that both agents in error cases and agents in success cases seem rational in their actions/attitudes.

Achievement Corollary 1: The agent reacts for the normative reason that p.

Achievement Corollary 2: The reason that p explains why the agent φ-s.

However, things don't always go this smoothly. Agents often act in the light of considerations which appear to them to be reasons for their actions without being in the conditions of the Achievement case.

First, agents may be guided by facts which they take to have normative bearing on some action, but which do not bear normatively on that action. This may be true of, for example, sexist agents, who believe someone's biological sex or gender constitutes a reason to deprive them of some resource or a reason to exclude them from some activity or job. More importantly, it may be true for agents who simply fail to recognise the normative bearing some fact has for their actions. In both types of case, agents are still reacting in the light of *facts*.

Here is an example, with the relevant corollaries added. The case is one of *believing* for normative reasons, but this does not matter here.

> ***Fact-Guidance (Fortunate Consequent-Affirmer).*** Sam wonders whether Terry took the bus to work. He knows that Terry's car is in the driveway. This is, in fact, a sufficient abductive reason to think that Terry took the bus. Sam also believes that if Terry took the bus, then Terry's car is in the driveway. But he comes to believe that Terry took the bus by inferring that he took the bus from his own belief that Terry's car is in the driveway and his belief that if Terry took the bus, then Terry's car is in the driveway by following an invalid deductive rule: from <if A then B>, and <B>, infer <A>. Sam hereby manifests a general consequent-affirming incompetence. (Lord and Sylvan 2019, 148) [93]

Fact-Guidance Corollary 1: The agent reacts on the basis of the fact that p.

Fact-Guidance Corollary 2: The fact that p explains why the agent believes that q.

Note that the consideration which constitutes a reason in Sam's case does appear in the two corollaries, just not qua normative reason. Just because Sam reacts in the light of facts without recognition or the wrong recognition of their normative bearing does not mean that we cannot explain Sam's reaction. Agents like Sam still seem to exercise a capacity to recognise truths and react on the basis of them. He is successful at least in

[93] Cases similar in structure with respect to our judgments of reasons-responsiveness already famously feature in Kant's discussion on acting on the motive of duty. I discussed versions of them in chapter 3 when touching on the concept of moral worth.

this respect. Hence, we can still explain Sam's reaction in terms of a fact (and his capacity to recognise such facts). Lord (2018,152) calls these explanations 'rationale explanations' to distinguish them from explanations involving normative achievement, but I shall simply refer to them as *fact explanations* (a more neutral notion, it seems to me).

Before I discuss the next batch of errors, I need to engage in a dialectically important distraction. We may refer to another familiar use of the concept of a reason to describe the difference between Achievement and Fact-Guidance. The distinction is sometimes captured by pointing out that Sam is still reacting for *motivating reasons*.

This terminology is notoriously tricky however, so I'll have to say a bit more. The term 'motivating reason' is used to refer to the consideration in the light of which the agent acts, that is, the consideration which, from her perspective made her reaction intelligible by casting it in a favourable light.[94] I take this to be equivalent to the usage "reason for which someone reacts" (Alvarez 2010, 36; Dancy 2000, 1). The same concept has also been expressed by "operative reason" (Scanlon 1998, 56) and "the agent's reason" (Enoch 2011). 'Motivating reason', as it is used here, is to be contrasted with 'explanatory reason', following Alvarez (2010). To see the difference, consider the well-trodden example of Othello whose jealousy leads him to kill Desdemona. There is an explanation available here according to which Othello killed Desdemona because he was jealous. But this is not an explanation that reveals what drove Othello, from his perspective, to kill Desdemona. We can imagine an Othello who is utterly bewildered by his own actions. It would still be true that there is an explanation in terms of jealousy for this Othello, but it would make little sense to say that he was reacting on the basis of (motivating) reasons. This is because Othello's reason for killing Desdemona is the putative fact that she was unfaithful to him. It is this consideration that figures in Othello's reasoning and imbues his action with intelligibility. The most important and essential feature of motivating reasons is then that they make an agent's action intelligible from their perspective. This feature is also why I am not going to engage with theories of motivating reasons that claim they are mental states. Mental states may be explanatory reasons (like Othello's jealousy), but the corresponding motivating reasons must be the contents of those mental states. The state itself becomes transparent when we are trying to pinpoint the agent's reason for reacting.

[94] There is another usage of 'motivating reason', on occasion presented as competitor to the factualist usage, which interprets motivating reasons as psychological states (Davidson 1963). I will not engage with this conception here because it does not overlap with the phenomenon I am tracking. Responsiveness to normative reasons means responsiveness to normative truths. When agents act on these truths, what explains their reactions are the truths themselves, not the mental states in which agents recognise them. Recognitional states become transparent in reasoning and reacting for reasons, as it were. For arguments against psychologism (including the transparency argument), see: Alvarez (2008, 2010); Dancy (2000, 2014); Hornsby (2008); Hyman (1999, 2015); Littlejohn (2012, 2014a); Raz (1999) and Williamson (2000).

As will hopefully become clear once I present my account of how these cases are related in terms of responsiveness to reasons, I am somewhat uncomfortable with the terminology of 'motivating reasons'. I don't think it cuts across cases in a helpful way. As I already said, I think the decisive distinction we ought to track in these cases is whether or not the capacity to respond to reasons was exercised on the relevant occasion. But as I shall argue later, acting in the light of what appears to the agent to be a reason may be true both of agents who haven't exercised their capacities and agents who have. The term 'motivating reason' from this perspective is unhelpful because it fits two essentially distinct types of explanations: p explaining why the agent reacted in terms of an exercise (of RR) and p explaining why the agent reacted, but not in terms of an exercise (of RR). I will therefore try to avoid the motivating reasons terminology where possible (it won't be possible consistently unfortunately because the terminology dominates the relevant literature).

End of distraction. Let us get back to the last category of error that is important. So far, we have considered agents who act in the light of truths. But agents may simply be wrong about the putative facts they take to be reasons (they may also be *accidentally* right, of course). In these cases, an agent makes a factual mistake, but they still treat the relevant proposition as a reason.

Such agents are perhaps most tricky to describe without presupposing a position in the debate. Let me say this. At the very least, it should be admitted that no fact explanation is true of these agents, because there is no fact to explain their reactions. However, it should also be admitted that it is not as though *no* explanation is forthcoming. After all, when agents react under the impression that some proposition is a reason for them to react, we will in hindsight be able to refer to their impression in order to provide a rationalising explanation. Why did I sprint to room 1.03? Because the lecture took place there, which is what I thought was the case (but turned out be false). Surely this constitutes a rationalising explanation which at very least *involves* the false proposition that the lecture took place in room 1.03. That is, surely we don't want to suddenly hold that what explains my sprinting is my belief rather than what it is I believe. For it is still true in this case that as I am sprinting, the consideration in light of which I act is that the lecture is in room 1.03, not the consideration that I believe the lecture to take place there.[95] Hence, when agents act on the basis of falsehoods, we may say they react for considerations which they believe to be true and treat as reasons. Consequently, there

[95] The trickiness of the case lies in the question whether false propositions are 'part of the world' in the same way that facts and normative reasons are (Alvarez 2010, ch. 5.5; Dancy 2000, 115). But it seems to me that there is a lot of metaphoricism involved in raising this concern. If we are unphased by the assumption of this thesis that normative facts are 'out there' in a mind-independent sense, then I don't really see why we would be worried by the concession that falsehoods are in a similar enough sense, out there as well.

is an explanation available for their reactions in terms of what I shall call *believed-to-be-true considerations*[96] to distinguish the relevant explanations from those that involve facts and reasons. I hope that this is a sufficiently neutral way to describe the explanations we offer for reactions under factual error.

The terminology of motivating reasons is already causing trouble at this stage. Some would hold that no proper reasons-vocabulary should be employed for such agents (Alvarez 2010, Littlejohn 2012, Hyman 2015). For them, these agents act for no (real) reason at all (at best, their reasons are counterfeit). Others, like Dancy (2000) and Lord (2018), would concede that these agents still act for motivating reasons and that their actions still have the same type of rationalising explanation as the agents in Fact-Guidance do. We will get to these positions later (they're both unsatisfactory).

Two variations of reactions based on falsehoods will be important. The first type of case is more philosophical in nature.

> **No Mistake Consideration-Guidance**. Boris is just like Rita from Achievement in every relevant respect and he finds himself in the very same situation – the smoke, the heat, the fire alarm. Boris too is an epistemically attuned and conscientious person in this situation, so he too takes an appropriate amount of time to check everything. Nothing appears out of order, so Boris believes the hotel is on fire and leaves the hotel based on this consideration. Or so it appears. For Boris is in a bad case. The hotel is not actually on fire, all indicators in favour of this assumption being manufactured by the fire department in an elaborate fire drill.

Consideration-Guidance Corollary 1:	The agent acts for believed-to-be-true consideration that p.
Consideration-Guidance Corollary 2:	The believed-to-be-true (btbt) consideration p explains why the agent φ-ed.[97]

Some might believe that the assumption that Boris thoroughly checked his evidence is incompatible with the assumption that the fire department managed to deceive him. Boris must have missed some obvious undercutting evidence, like two guys from the fire department powering a smoke machine. It is this kind of philosophical lack of

[96] This formulation intentionally resembles what Dancy calls an 'apositional account' (Dancy 2000, 128), according to which what explains the agent's action is *what they believe*, although their believing it figures as an important background condition.

[97] It can't be the fact that he believed it to be true because this is not what explains Boris's action from his perspective.

imagination, it seems to me, which has driven philosophers to replace the fire department with an evil, all-powerful demon, whose powers of deception outstrip any of the epistemic powers Boris could muster. If you must, assume such a demon is deceiving Boris. Since Boris makes no relevant mistake in his case, let us say that Boris is in the No Mistake Consideration-Guidance case.

I should mention that I will not engage with the recent idea that agents in New Evil Demon scenarios (scenarios like the one Boris is in) *do* have a set of reasons-giving facts at their disposal after all, namely a set of evidential facts, most important among them facts about how things *appear* to the agent (see Lord 2018, ch. 7.4-7.6; Williamson 2000, 198; Kiesewetter 2017, ch.7.5-7.6). As far as I am concerned here, there are no normative reasons for Boris to react to, not even so-called 'backup reasons'.[98]

The case of Boris, a diligent but unfortunate fellow, is perhaps a philosophical idealisation. Often when agents fail to latch onto the facts that provide their reasons, they have made a mistake. The mistake might often be tiny and hard to notice, but a mistake it will be, nonetheless. Here is an example:

> **Tiny Mistake Consideration-Guidance**. Hannah is in the position of Boris with the notable difference that she has access to the information that would reveal the machinations of the fire department to her. Hannah was told by the concierge that a fire drill is coming up earlier, a fact which she has since failed to properly keep in mind. This does not mean that Rita has strictly speaking forgotten the information, she just fails to connect it with her current situation. Believing her hotel to be on fire, Hannah leaves the hotel.

It should be clear that the Consideration-Guidance Corollaries are true of Hannah as well. She too acts on the basis of believed-to-be-true considerations, and these considerations afford us a rational explanation of her action.

With this, we have the spectrum of cases that I believe are central on the table. The question I shall ask now is how we should arrange these cases in terms of whether and to what extent they contain responsive agency. Two principled answers to this question are available. According to the first, the Univocal answer, these cases share significant similarities, and the relations involved in them should therefore be understood in terms

[98] I believe the back-up view significantly distorts the epistemic role appearances play for agents. When I leave the hotel because it appears to me that it is on fire, then the consideration in light of which I act is not the fact that it appears to me that the hotel is on fire but the content of my appearance – i.e. that the hotel is on fire. I act in the light of appearance-facts only when it has already been pointed out to me that I might be in non-veridical circumstances (but I don't want to risk burning to death anyway). To my knowledge, only Lord (2018), ch. 7.4-7.6 properly acknowledges this problem. I don't have time to engage his arguments on non-inferential basing in that chapter (I find them unconvincing).

of these similarities. According to the second, the Non-Univocal answer, the success cases should be understood as an entirely different species.

3. Univocal vs. Non-Univocal Views of Responding to Reasons.

Notice that at first glance, there is a through-line for the above cases. This is because it seems like Rita from the Achievement scenario has, besides the achievement explanation in terms of her normative reasons, also a fact explanation and a believed-to-be-true explanation available for her. That is, the Fact-Guidance and the Consideration-Guidance corollaries seem true of Rita in the Achievement scenario as well. After all, when Rita is guided by a reason, then she is also guided by the fact which provides this reason and when Rita is guided by her reason, she is also guided by a believed-to-be-true proposition (a proposition, which, for her case, *is* true). Thus, we can say that Rita does not only react for normative reasons, she also reacts on the basis of facts, and believed-to-be-true considerations.

Further, the Consideration-Guidance corollaries seem true of Sam from the Fact-Guidance scenario as well. After all, when Sam acts guided by a fact, he is also acting for a believed-to-be-true consideration (a consideration, which, for him, is true). So Sam's belief cannot only be explained in terms of a fact, but also in terms of a believed-to-be-true consideration.

If we focus on these similarities, it will seem like there is a fundamental corollary here, which is the Consideration-Guidance corollary (both corollaries, that is). Since the relations expressed in this scheme seem to hold for the agents in all the others, it will seem prima facie appealing that they are the fundamental ones, from which all others can be built. Perhaps, that is, agents who act on the basis of facts as in Fact-Guidance scenario are really just guided by believed-to-be-true considerations that *are* true, and agents who act for normative reasons are really just acting on the basis of facts that happen to be normative reasons. Perhaps there is only one way to relate to reasons and it is the way that agents in Consideration-Guidance relate to what they believe. Agents in Fact-Guidance and Achievement will then relate to considerations that are in addition true and have normative bearing.

Although ideas like this are plentiful in philosophy, they are rarely explicit. Here is one exception:

> Since believing for a good reason is believing for a reason (one that is good), the account [of reacting for a reason] clarifies believing for a good reason. […] Indeed, if an indirectly (prima facie) justified belief is simply a belief held for at

> least one good reason, then if our conditions [for reacting for a reason] are supplemented with an account of what constitutes a good reason, we shall have all the materials we need to understand one of the main kinds of justified belief and, in good part, one of the main kinds of knowledge. (Audi 1993, 267)

Reacting (Audi is talking about beliefs, but the idea generalises) for a good reason for Audi just is the same relation as reacting for a reason, which he thinks is the underlying phenomenon. It is reacting for a reason that is good.

When we think of our cases, there is some strong prima facie motivation to think that they are all instances of responsive agency. More precisely, we can glean from them the idea that there is a Univocal[99] analysis of responding to reasons.[100] Responding to facts, responding to believed-to-be-true reasons, and responding to normative reasons are all instances of responsiveness, according to such an analysis, they are cut from the same cloth. Correspondingly, the underlying explanatory relation is also the same, which vindicates Williams famous dictum[101] that "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 1979, 102)

This is then the Univocal View:

> **Univocal View**: There is only one relation[102] involved in reacting for normative reasons, reacting for facts and reacting for believed-to-be-true considerations (See Lord 2018, 151).

The Univocal View has seen some pushback lately. The opposing viewpoint is that some of the relations involved in the above cases are distinct. Most obviously, as Lord argues, we can think that reacting for normative reasons is a distinct achievement relation, in which a normative consideration in the world explains our behaviour through our abilities, whereas acting for mere facts or believed-to-be-true considerations does not involve such achievement and essential connection to the normative world. I have pointed out before that this view runs parallel to those views that hold that knowledge is

[99] I follow Lord (2018) in this label.

[100] Notably, this factoring picture is just a strong motivation for Univocal Views. Schroeder (2008) rejects the picture, but still argues that recognising reasons has a univocal analysis. Perhaps even more confusing Lord (2010), whom I here presented as the champion of the alternative (Non-Univocal) seems to defend the factoring picture. However, Lord's defence consists in pointing out reasons that agents in error cases are in successfully in contact with after all, so the base dialectics doesn't change. This is part of a strategy I cannot discuss in the thesis, but which I briefly address in section 2, esp. footnote 9.

[101] See also Dancy (2000),121; Dancy (2008), 267-268.

[102] The relation is between agent, reason, and action. I.e. the reason explains the agent's action.

a special distinct state – a 'broad' mental state some say[103], which cannot be constructed from two (or more) separate conditions, one internal one external. If this is our line of thought, we will think that responding to reasons is radically split. There is on the one hand reacting for normative reasons, an achievement relation the worldly component of which cannot be detached from the concept. There is on the other hand reacting for facts which happen to be reasons and reacting for believed-to-be-true considerations, which only offer weaker rationalising explanations.

Since there is some debate over whether reacting on the basis of believed-to-be-true considerations should count as reacting for reasons at all, there is more than one way to hold this position. Lord, who thinks that both agents in Fact-Guidance and Consideration Guidance are related to their 'motivating reasons' (Lord 2018, 150), sees the decisive split as that between reacting for normative and reacting for motivating reasons. Others (Alvarez 2010, Littlejohn 2012), who deny that acting on the basis of false propositions is reacting for any reasons (not even motivating reasons), will see responsiveness as split in three parts: genuine achievement (acting for normative reasons), Guidance by facts (acting for motivating reasons), and acting for believed-to-be-true considerations.[104] No matter how exactly we partition the scenarios of Achievement, Fact-Guidance, and Consideration-Guidance, though, the Non-Univocal idea will be that we cannot synthesise the relations in one of them by using the others.

> **Non-Univocal View**: We stand in a different relation to normative reasons when we react for normative reasons than we do to considerations when we react on the basis of facts, and when we react for believed-to-be-true considerations.[105] (Lord 2018, 152)

I want to assess the central argument against the Univocal View next. Before I do, let me briefly address a dismissive attitude I have sometimes encountered with respect to the question whether responding is a unified phenomenon.

The attitude is that it must be a kind of verbal question whether responding is unified or not. In some sense, surely, everyone must agree that reacting for normative reasons and

[103] As far as I am aware Williamson does not prominently use this terminology, preferring to argue that knowledge is a mental state which cannot be understood conjunctively (Williamson 2000, 48), but others, such as Fricker (2009) for better or worse use the 'broad' terminology.

[104] It is possible of course to hold that Fact-Guidance and Achievement involve the same relation, while Consideration-Guidance does not. But I would think that usually when we are convinced that there is an important difference between Consideration-Guidance and Fact-Guidance, we will also hold that there is an important difference between Achievement and Fact-Guidance

[105] Notice that Disjunctivism is compatible with thinking that acting for normative reasons implies acting for 'motivating reasons'(in Lord's sense). The important thing is that these notions still require separate analyses because they are models for two separate phenomena.

reacting for facts or believed-to-be-true considerations are both instances of responsive agency, this attitude holds.

It is true that it isn't as clear as possible in this debate in what sense the explanations that are available in Achievement, Fact-Guidance, and Consideration-Guidance are the *same*, or not the same (see Cunningham 2019b for a discussion). But this should not stop us from recognising that there is an intelligible sense in which they are different if the Non-Univocal View is true. The sense in which different relations are involved is the same sense in which agents who know are related differently to the world than agents who merely truly believe, if it is true that knowledge isn't just belief plus the relevant success. If we find the latter distinction intelligible, we must find the discussion between Univocal and Non-Univocal Views intelligible. For both involve the very same idea: the idea that the orthonomous relationship we have with the world in cases of achievement is of a different breed entirely that the relationship we have with the world in cases of failure or accidental success.

With this out of the way, I will now show how the core argument for the Non-Univocal View falters when faced with the results from the previous chapter.

4. Lord's Core Argument Against the Univocal View

Lord (ch.6) raises a variety of problems against the Univocal View, but I take it his core complaint is directed against one of the major motivations of the Univocal View. Recall that I introduced the view by pointing out that it seems like the relations that hold in the Fact-Guidance, and in the Consideration-Guidance scenario *also* hold in Achievement scenario. There is therefore the temptation to think that there is just one relation to reasons involved in these cases, from which the others can be constructed by adding separate conditions. I sometimes referred to this idea as the factoring picture, but another way to express it is to say that since it is assumed that there is only one underlying relation in all three scenarios, the other relations are *composites* of the core relation and some additional element – an element of external success most obviously. Let me, in reconstructing Lord's argument, adopt his parlance, according to which agents react on the basis of motivating reasons both in Fact-Guidance and Consideration-Guidance scenarios (so acting on the basis of false considerations can count as acting for motivating reasons). I will tease the two apart again below.

The idea that reacting for normative reasons as in Achievement is a composite notion can then be expressed as follows:

> **Composite**: S reacts for the normative reason p iff p is S's motivating reason to φ and p is a normative reason to φ.[106]

Recall that this is precisely the idea that Audi expresses in the quote above when he alleges that to act for a good reason just is to act for a reason – one that is good.

The problem with such approaches to any orthonomy notion is that they are vulnerable to cases of accidentality in the form of cases of deviance. Recall the following agent from the Fact-Guidance scenario.

> **Fortunate Consequent-Affirmer**. Sam wonders whether Terry took the bus to work. He knows that Terry's car is in the driveway. This is, in fact, a sufficient abductive reason to think that Terry took the bus. Sam also believes that if Terry took the bus, then Terry's car is in the driveway. But he comes to believe that Terry took the bus by inferring that he took the bus from his own belief that Terry's car is in the driveway and his belief that if Terry took the bus, then Terry's car is in the driveway by following an invalid deductive rule: from <if A then B>, and <B>, infer <A>. Sam hereby manifests a general consequent-affirming incompetence. (Lord and Sylvan 2019, 148)

Why do I call this a deviance scenario? Because just like in the cases briefly reviewed in chapter 3 (and 2), the proposition that provides (or constitutes) a reason for the agent causes (and in fact also guides) the action and yet it seems like an accident that Sam believes what he has reason to believe. And this is the problem for Composite. For if it is an accident that Sam believes what he has reason to believe, then we certainly shouldn't say that Sam believes *for the reason* that Terry's car is in the driveway. But Composite is satisfied in Sam's case. The reason which motivates Sam's belief is also the reason that justifies it. Sam believes for a reason, one that is good, in Audi's terms. Still, he doesn't believe for a normative reason.[107]

What we have in Fortunate Consequent Affirmer, it seems, is a failure of a composite condition. Reacting for reasons cannot just be reacting for motivating reasons that also happen to be normative reasons. Evidently this puts in jeopardy the factoring picture that is the main motivation for a Univocal View.[108]

---

[106] As I have alluded to in several places already, these considerations about compositeness bear similarities to Williamson's (2000) arguments for the idea that knowledge is a prime notion, i.e. not decomposable into a belief component and a success component which can be understood independently. However, while Williamsons cases do clearly invoke intuitions about non-accidentality, he does not seem to be aware of this, nor does he develop them in a way that undermines simple causalist approaches (see Brueckner 2002 for this criticism).

[107] If you have doubts about the claim that Sam doesn't believe for a reason, notice that the reason doesn't offer the type of achievement explanation it would if Sam was in an Achievement type case.

[108] The next chapter will go into much more detail about why these accidentality cases exist.

I briefly explored in the last section of the previous chapter the thought that the missing element to get a real achievement in deviance cases is the exercise of the relevant capacity. The idea applies to the current deviance case as well. In being guided by the fact that Terry's car is in the driveway, Sam doesn't manifest his capacity to react for normative reasons (instead some strange dispositional system is active), according to this idea. Sam isn't reacting for a reason then, because his success wasn't brought about by the involvement of his capacities.

Lord suggests the same basic solution (Lord 2018, ch. 5.4.2). However, it is exactly at this point where I believe Lord's argument falters by failing to consider an important logical option. Lord treats the cases that show that we need an exercise condition as a non-accidentality clause as ipso facto also falsifying any Univocal View. Inversely, he treats Univocal Views as automatically committed to a composite condition of some sort:

> (…) we are now in a position to see that there are much more general reasons to reject Univocal Views. These reasons mirror the reasons to reject univocal views in the philosophy of perception. Univocal views about perception and reacting for reasons fail to explain why the good cases involve achievements. This is because it's not plausible that the relation involved in cases of achievement can be built out of the relation involved in the cases that don't involve achievements. (Lord 2018, 169)

When Lord refers to "building one relation out of the other", he is referring to the failure of the composite condition discussed above. The thought is that the relation in cases that involve achievements cannot be compositely synthesized from cases that don't involve achievements because achievements require the exercise of a capacity.

However, here a crucial premise is missing. For why could a Univocal View not simply take this lesson on board, claiming that both reacting for normative reasons and reacting for motivating reasons involve the exercise of a capacity – namely the capacity to respond to reasons? The only reason against such a move I can see is that Lord thinks something about exercising a capacity itself is incompatible with the central commitment of the Univocal View to treat reacting for normative reasons and reacting for motivating reasons as interestingly unified. So what Lord needs is a reason why agents who react on the basis of false propositions cannot also exercise the capacity to respond to reasons. And the only assumption that would provide such a reason is the assumption that *a reaction is the exercise of the capacity to respond to reasons only if it is a successful response to reasons.* If this assumption was true, it would follow that agents in error cases are not exercising the capacity to respond to reasons and consequently that no Univocal

View is entitled to using the notion of the exercise of a capacity. That Lord implicitly holds this assumption becomes clear earlier in his book, where he asserts:

> On the natural interpretation of Normal Colloquium [Achievement], I act for the normative reason provided by the fact that the colloquium starts at 4 pm. (…) I do that when my action is a manifestation of a disposition sensitive to the property of being a normative reason to go to Robertson Hall. I cannot be manifesting such a disposition in Unusual Colloquium [Consideration-Guidance]. This is because there is no fact that colloquium starts at 4 pm that has the property of being a normative reason to go to Robertson Hall. So I cannot be manifesting an essentially normative disposition with that as its manifestation condition. (Lord 2018, 151, my inserts)

In order for it to follow that I am not exercising a normative disposition in acting on a false proposition, it has to be true that exercising is 'factive', i.e. that whenever I exercise a capacity, I exercise it successfully. But this assumption is false. I argued against it in length in the last chapter. For it is just a version of what I there called (Success).

This then finally allows us to see the whole argument in the background of Lord's Non-Univocal View. The argument can be summarized like this:

(1) Reacting for normative reasons is not a composite (it is an achievement).
(2) Non-composite notions (achievements) require the exercise of a capacity (from Sam's Case).
(3) Exercising is factive (Success).
(4) Univocal Views require non-factive relations.
(5) So Univocal Views cannot capture the fact that reacting for normative reasons is an achievement.

I deny (3). More precisely: I agree that the exercise of capacities is needed to exclude the types of accidentality cases that lead to a non-composite picture. But I disagree that the feature of the notion of an exercise responsible for facilitating such non-accidentality is its factivity. Chapter 7 is dedicated to understanding better why exercises block accidentality and how they establish non-composite relations. Here, I will just develop a first attempt at expressing the idea: the underlying explanatory relation in all cases is one which involves the exercise of a capacity.

Before I elaborate my account, let me address something I ignored earlier for the sake of lean presentation. Technically, Fortunate Consequent Affirmer is a case in which we have available a fact explanation in terms of a fact that is also a reason but no achievement explanation. This means that we cannot understand the relations involved

in Achievement in virtue of the relations involved in Fact-Guidance. Now I have explained in the case in a way that does not require the problematic concept of motivating reasons. In fact, we can see more clearly now just in what way it is doing problematic work in Lord's argument. As I already pointed out, Lord takes both the agents in Fact-Guidance and the agents in Consideration-Guidance as acting for motivating reasons. His adherence to the thesis that exercising is a success notion then leads to the conclusion that cases of Consideration-Guidance also cannot be involved in the achievement relation in Achievement. But my denial of (3) above already allows us to notice something interesting here. Although Sam's belief is explained by a fact and Rita and Boris reactions are not, the reactions of Rita's and Boris's are not accidental in any way. In this sense, Rita and Boris appear much closer to being responsive to reasons than Sam is. In slogan form, what my argument allows us to see is that accidental success is worse than non-accidental failure. The agents in the Consideration-Guidance scenarios exhibit high attunedness to the normative aspects of their situation, although they unfortunately either make a mistake or are prevented from correctly interpreting their environment by deception. The agent in the Fact-Guidance scenario is completely off in tracking normative importance, so we should class them as not responsive. Factual errors don't weigh against judgments of responsiveness as heavily as do normative errors. But Lord uses the concept of motivating reasons both for agents under factual error and agents under accidental success, which suggests that they are on a par in terms of responsiveness. As I shall suggest in more detail below, we should think of agents like Rita and Boris as exercising the capacity to respond to reasons and hence as responding to reasons. But we should think of agents like Sam as failing to exercise their capacity and hence as not responding to reasons. Claiming both types of agents react for motivating reasons obscures this essential distinction.

Let me elaborate on this account now.

5. The Exercise Univocal View

5.1. The Core View

In accordance with what I have just discussed, I propose that the crucial aspect according to which we measure agents' responsiveness in Achievement, Fact-Guidance, and Consideration Guidance is not whether they react to motivating reasons, nor is it whether all of their reactions can be explained in terms of believed-to-be-true considerations. Instead, the distinctive factor is whether the agent is still exercising their capacity to respond to reasons. Since we know that Sam from Fact-Guidance doesn't exercise that capacity (for his success is accidental) and we know that Boris and Rita do (because their

failure is evaluable according to the exercise standards of the capacity to respond to reasons), my view proposes that the spectrum is split between Rita, Boris on one side and Sam (and other even more off track agents) on the other.

This is the structure of the view I am proposing. Let me now fill in this structure a bit more.

The basic notion of my view is that of an exercise of the capacity to respond to reasons. When some φ-ing counts as response to a reason, then the φ-ing and the reason count as related through the exercise of the relevant capacity (the nature of this exercise relation is explored in chapter 7).

Recall that my hypothesis in the previous chapters was:

> **Responding**: To respond to a reason is to exercise the capacity to respond to this (highly) specific type[109] of reason.

I stated in the beginning of the chapter that I will rely here on intuitive judgments on what it is to exercise a capacity. I want to extend this approach here just a little bit. In keeping my focus on the explanatory relations that go with the scenarios discussed above, I want to point out that exercising capacities comes with an explanatory corollary as well.

Part of what we are saying when we say that an agent exercised a capacity is that we can explain their action by describing it as the exercise of that capacity. The capacity in these cases is involved in understanding why the agent acted (or formed an attitude). Now, of course it will in these cases not be the capacity of the agent *alone* that does the explaining. After all, normative reasons, facts or believed-to-be-true considerations are involved as well. But to say that agent exercised a capacity in reacting to a reason is to say that *part* of the relevant explanation is the agent's capacity. We can see that there is such a difference by looking at Sam the Fortunate Consequent Affirmer again. Sam's reaction is guided by facts (and presumably a basic capacity to be guided by facts), but what is *un*involved in the explanation is the capacity to respond to reasons.

I have chosen to express the connection between φ-ing, the reason for φ-ing, and the exercise of the capacity to respond to reasons by saying that φ-ing is explained by p "in terms of" φ-ing being an exercise. Right now, this just means that p explains φ-ing in what I shall call an *exercise-explanation* and "in terms of" is my grammatical vehicle for saying this. But the grammar here does point to my eventual account of exercise-explanations: they involve three relata, the reason, the φ-ing and the capacity "itself" in a sense.

[109] Recall that I hold that the RR capacities relevant for orthonomy are highly fine-grained.

I take it that in this abstract form, this implication of the notion of exercising a capacity should not be very controversial. The third part of this thesis is dedicated to spelling out the explanatory role of the dispositional property (ch.7) – of which I assume capacities are a subclass. In short, I will argue there that when O exercises the disposition d to φ in C (triggering circumstances), d explains why C caused O to φ.[110] For now, I think we can safely rely on the intuitive idea of an exercise-explanation – an explanation in which the exercise of a capacity plays an indispensable role.

Exercise-Explanation: q explains p in terms of an exercise of d only if the exercise of d is indispensable for the truth of 'p because q'.

The basis of my account is then to point out that both reacting for believed-to-be-true considerations and reacting for normative reasons are instances of responding to reasons, because both involve the same relation: that of exercising the capacity to respond to (sufficient or decisive) normative reasons. This means that the same explanatory relation is involved, that which proceeds in terms of the exercise of the capacity to respond to normative reasons. In accordance with the results of the previous chapter, this is a capacity to correctly pick up and translate into action real normative significance.

The difference between reacting for believed-to-be-true (btbt) considerations and reacting for normative reasons is then relatively unimportant. In the case of reacting for normative reasons, the reason that p explains the agent's reaction in terms of their capacity to respond to reasons. In the case of reacting for believed-to-be-true considerations, the fact or false proposition that p explains the agent's reaction in terms of their capacity to respond to reasons. To keep this in mind:

Reacting for normative reasons: An agent S φ-s for the normative reason r just in case S's φ-ing is explained by r in terms of φ-ing being the exercise of the capacity to respond to reasons.

Reacting for btbt-considerations: An agent S φ-s for the btbt-consideration that p just in case S's φ-ing is explained by p in terms of φ-ing being the exercise of the capacity to respond to reasons.

[110] Metaphysically, dispositional properties serve as sustainers of causal processes in which their triggers are involved, on my preferred view. But I lack the room here to develop this proposal.

This distinction is more helpful than the motivating reasons terminology. Motivating reasons come cheap, because even quite crazy agents may treat a consideration which they believe to be true as a reason (more on this below). What we need to treat as a unified phenomenon are cases in which the agent's motivation is driven/produced by their rational capacities. If we need to, we can express this difference by introducing two types of motivating reasons relations: those in which a proposition (true or false) explains a reaction in terms of the exercise of a RR capacity, and those in which a proposition (true or false) explains a reaction without such involvement of a capacity. We can call the former motivating considerations *rational* motivating reasons and the latter *arational* motivating reasons. But then again, we can just stick with my primary terminology.

Right now, let me go through how my Univocal View handles the differences and similarities between Achievement, Fact-Guidance, and Consideration-Guidance in detail.

Let us start with Sam, the fortunate consequent affirmer from the Fact-Guidance scenario. Recall that Sam ends up believing what he has reason to believe being guided by the very fact which provides this reason. But Sam still believes what he has reason to believe accidentally. Hence, he does not respond to the reason that Terry's car is in the driveway (even though he does recognise it as a fact).

Here, my account can agree with Lord: The reason Sam is unresponsive is that he fails to exercise his capacity to respond to reasons. This is why he occupies the very fringes of the spectrum of responsiveness. Notably, my account opens up the possibility for treating the case in a more nuanced manner than it is treated by Lord. Recall that Lord's original point was that even though Sam belief is guided by the fact that Terry's car is in the driveway, it is not guided by the normative significance of the that fact. Now, Lord assumes that we can still sufficiently understand Sam's belief so that we can say he reacted for motivating reasons.

It is of course true that we can explain Sam's belief in terms of providing some intelligibility of it by pointing to the fact that Terry's car is in the driveway and the consequent affirming rule that Sam is following. But this type of intelligibility explanation is available for even the most crazy and deluded agents as well. We can think of agents who are guided (in the minimal sense) by the fact that a child is drowning in the lake to eventually save it, but who act on the basis of rules entirely alien to us ('water on drowning children will doom the earth'). The same sort of intelligibility explanation will be true of these agents, but they are at the furthest distance from responsive agency we can imagine. Part of the confusion is perhaps caused by imagining Sam as making a minor mistake in his reasoning but securing external success nonetheless. But the way

Sam is described, he is manifesting a general consequent affirming disposition. Sam is therefore not reacting at all on the basis of any responsive dispositions of his.

So Lord's usage of the term 'motivating reasons' here is distorting the situation. Sam is reacting for arational motivating reasons, i.e. we can explain his reaction on the basis of a fact, but his rational capacities are uninvolved in such an explanation. He is not like Hannah and Boris from the Consideration-Guidance scenario, who make no or almost no mistake and get things wrong factually, i.e. whose RR capacities are involved in the explanation of their reactions.

Another way to frame this is to anticipate a bit of vocabulary that I shall introduce when I talk about the recognitional part of the capacity to respond to reasons. There, I will apply the same basic thought developed here to recognising reasons. I will say that agents who have a recognitional stance towards reasons present facts *under the guise of reasons* (Gregory 2013, Singh 2019). However, the mere guise of reasons comes cheaply. The crazy agents mentioned above probably act under the guise of reasons. What is important is that the presentation of a proposition under the guise of reasons is itself the exercise of the (sub) capacity to recognise reasons. The distinction maps exactly onto the difference between Hannah and Sam (and his crazy kin). Both act under a kind of normative guise, but only Hannah's normative guise is the manifestation of the capacity to respond to reasons. Therefore, there is an exercise-explanation, and thereby the possibility of criticism of a certain kind, available for Hannah but not for Sam and his kin. This seems to be exactly the kind of result we should aim for. See section 5.2. below for details on acting under the guise of reasons.

This is the segue to the situation of Boris and Hannah from the Consideration-Guidance scenario then. Both agents, it seems to me, are located much closer to Rita from the Achievement scenario than Sam is. Both agents *almost* get it right. Both agents and reactions seem rational moreover, with Boris being as rational as Rita in the Achievement scenario. Intuitively then, we should think that Rita (Achievement), Boris (No Mistake, but factual error), and Hannah (Tiny Mistake and factual error) cluster together, while Sam – and all those agents even worse than Sam – resides at a considerable distance in terms of reasons-responsiveness from Rita, Boris, and Hannah.

My account captures this intuition. Rita, Boris, and Hannah are all responding to reasons. Thus, an explanation in terms of the exercise of their reasons-responsiveness capacity is available for each of them. Hannah is exercising that capacity badly (though not *very* badly), so she has flouted some internal standard, as well as failing to live up to the external standard of the capacity. But she is still clearly reasons-responsive. We can

explain, with recourse to the exercise of her capacity to respond to reasons, why she didn't live up to the external success standard.

Boris is exercising his capacity to respond to reasons perfectly. He is like the tachometer I discussed in the previous chapter. His capacity is functioning well, but it lacks the opportunity to latch onto the real normative situation. The lack of opportunity does not mean that the capacity to respond to reasons isn't exercised, however. Sometimes despite the fact that our capacity to be in tune with the normative properties of our situation is on full display, the situation is setup in such a way as to make it impossible for us to correctly respond to those properties. In these cases, we are not only excused for reacting against the reasons there actually were, but might, if all other conditions are right, even be praised for it. An act of care in response to what the agent was – despite their best ability – deceived into thinking were good reasons to care will still be received with the corresponding evaluation. A false belief in response to what looked to the agent like excellent reasons to believe might still garner the relevant epistemic praise. It is for the reasons provided by cases like these that Boris's case is especially important. For his case shows that the external standards which are supposed to make cases like Achievement special intuitively fail to make an evaluative difference. Boris is as creditworthy as Rita is for his reaction. My account holds that both are equally creditworthy because both exercise their capacities and exercise them in accordance with all relevant internal standards for these capacities. Does this make Rita (Achievement) too similar to Boris (No Mistake Consideration Guidance)? One criticism of the Univocal View might be that assimilating the two makes Rita's success seem meaningless. Why go for the external success if all relevant evaluative standards can be satisfied internally, the thought might go.

In a sense, I think we do want Rita's success to mean less than it does according to the Non-Univocal View. After all, there is no evaluative practice, it seems, that we would reserve for Rita alone and not extend to Boris. That said, my account does allow us to say that Rita's success is not just "tacked on" as it were. Rita reacts on the basis of a real normative reason, not by accident but in virtue of her capacity to respond to reasons. So an exercise-explanation is available for her reaction that crucially involves a reason, not a mere fact or a false proposition. However, on my view, the more essential part about responding to reasons is that an exercise-explanation is available for the relevant response. Whether this explanation is factive or not, does not play much of a role. [111]

[111] This is where the constructivist or constitutivist undertones of this chapter come to the surface. Lord and other advocates of the Non-Univocal view are driven to this view because they wat objective normative reasons to do the bulk of the work of explaining either all types of normativity or at least rationality. This means that if there is a type of normativity that is tied to explanations – as there clearly is in all the cases we have discussed and in all the cases of "worth" concepts - nothing other than the real thing, the real objective reasons, will be enough for these theorists. But I hope to have shown that they thereby have to bend our conception of

The realisation that achievement is not bound to success in the way Lord suggests then opens up a more nuanced account of how achievement relates to error and success.

Hannah makes a mistake; she clearly fails according to an internal standard as well as an external one. But in another sense, in the sense of exercising her capacities, her response can still be classified as an achievement. It is still at least a response – albeit a bad one. In fact, we can only really say that it is a bad response because it is a response, and so an exercise of her capacity, in the first place (this invokes my 'bad action' argument from section 6 of the previous chapter).

So Boris and Rita are so similar to one another because they are both responding perfectly, and while Hannah also responds (and is therefore categorically different from Sam), she is less like the two of them because she fails in exercising her capacity.

This is, then, my version of a Univocal View. Responding to reasons is indeed a unified notion. It is unified by the exercise of the capacity to respond to reasons. This view can handle the relevant differences between agents because it takes the exercising relation as fundamental. It can also handle the relevant similarities because it does not have too strict of a view on what it takes to earn an achievement in responding to reasons.

These considerations were mainly aimed at the reactive aspect of responding to reasons. But in some of my discussion, I have relied on the assumption that agents who act in the light of believed-to-be-true considerations, act under 'the guise of reasons'. This terminology concerns the recognitional aspect of responding to reasons and is best discussed under that heading. Hence, the next section discusses what it takes to recognise reasons within the system set up in this section.

5.2. Recognising Reasons

It is generally recognized that a necessary part of responding to reasons is to establish some kind of epistemic access or stance towards reasons. Different names have been advanced for this stance, but the most prominent ones are "having" (Schroeder 2008, Comesaña & McGrath 2014) or "possessing" (Lord 2018, ch. 3) reasons.[112]

responsive agency itself. My account suggests another way to go. I think that most of the normative work can be done by the notion of exercising a capacity itself. So the worry that "exercising a capacity" in a sense swamps (recall the swamping problem for epistemic value) reacting on the basis of a real reason because both Boris and Rita exercise their capacities perfectly but only Rita reacts for objective normative reasons is only a worry if you want objective normative reasons to carry all the weight in the "final metanormative theory". I don't want this. Instead, I think an exercise-based account of reducing normativity makes a lot more sense and the Boris and Rita cases show this.

[112] The 'having' or 'possession' vocabulary is also used in an ex ante way, according to which "S has a reason to φ' is equivalent to 'there is a reason for S to φ'. Of course this is not the usage I latch onto here, since I am

In fact, the scenarios I discussed above – Achievement, Fact-Guidance, and Consideration-Guidance – arguably elicit differing intuitions on the responsiveness status of agents' actions because they involve epistemic failures on their part. Since I had to focus on reacting for reasons first for presentational purposes, I will now complete my account with my take on the 'having' or 'possession' condition.

My account of the recognitional part of the capacity to respond to reasons will rely again on the dialectical position developed in section 4. That is, it will focus on the notion of an exercise while relying on the results of the previous chapter in assuming that there are defective exercises of capacities. I will call exercises of the recognitional subcapacity a 'recognitional stance' (so as not to unnecessarily trigger natural language intuitions about 'having' and 'possessing'):

Recognitional Stance: An agent S has a recognitional stance towards p just in case S represents p under the guise of reasons and this representation is a manifestation of the capacity to respond to reasons.

When p is true, we may then say that a reason is possessed, if we want to have some notion to designate the success cases. But again, the important implication of Recognitional Stance will be that whether an agent takes a recognitional stance towards a proposition does not depend on the truth of that proposition.

Let me now explain how the view fits into the debate over having reasons.

What is it to have a reason? It is to have some kind of epistemic connection to it, or less tendentiously, to have some kind of epistemic stance towards a proposition. This stance is relatively easy to spell out when the agent successfully picks up a normative reason. For example, when Rita sees that the hotel is on fire and subsequently represents that her hotel is on fire, she represents a fact that counts as a normative reason to leave the hotel. She thereby, one might think, *has* that reason. We already know that we need to slightly improve this condition. It is not enough that Rita represents the fact that the hotel is on fire and that the hotel is indeed on fire for her to have a reason in this case. For the connection between the fact and the representation might be purely accidental, as when Rita gets hit in the head and hence believes the hotel is on fire. So in order to have a reason, Rita needs to represent the fact that the hotel is on fire where this representation is a exercise of her capacity to respond to reasons.

interested in the agent's responses to reasons, not the deontic facts about what reasons there are in the first place.

Note that there are prima facie two ways to come to represent a reasons-fact: by direct, non-inferential exercise of the capacity to respond to reasons, that is, the subcapacity to pick up facts that constitute reasons, and by indirect, inferential exercise of that capacity. The hotel case is plausibly a case of non-inferential representation, if we think Rita just directly sees that the hotel is on fire. I will subsequently treat the case like this. But we could also think that Rita infers that the hotel is on fire from the indicators she picked up non-inferentially – i.e. the smoke, the heat, the smell etc. The details of the philosophy of perception we prefer here are unimportant. What is important is that we recognize that having reasons may involve the exercise of both inferential and non-inferential capacities.

It is plausible that representing (competently - I will omit this in the following discussion) the relevant facts, or at least facts suitably connected to those facts (see Mantel 2018, Ch.8 for a discussion of how "indicators" of facts are also sufficient) is a necessary condition for having a reason. But it isn't sufficient.

It isn't sufficient because factual error is not the only kind of error pertinent to reasons-responsiveness. To see this, consider Adina who correctly represents her hotel as being on fire, but who believes the fire will not hurt her due to the fact that a highly reliable expert on hotel security told her so. Adina correctly represents the fact that the hotel is on fire, but she fails to see the normative relevance of this fact. She fails to see that it favours leaving the hotel immediately. Adina represents the reasons-providing fact *de re*, we may say, and so we may call views of having reasons according to which it is sufficient to have a reason to present it de re *De Re Views.*[113]

But representing a fact that is a reason de re is not enough to 'have' that reason, as Adina's case shows. Somehow it must further be the case that the agent has this fact *qua* reason (Sylvan 2015 calls this the "unapparent reasons problem").

This triggers a familiar discussion about contents of mental states. The most obvious answer we could give to the question what exact condition is missing from the case where Rita doesn't pick up the normative relevance of a fact is that Rita's mental state doesn't represent in its content that the hotel being on fire is a reason for her to leave to hotel. This would amount to saying that in order to have (and subsequently react for) reasons, we need to have contents of the form 'the hotel being on fire is a reason to leave it'.

[113] De Re Views are held by Parfit (2001, 2011); Schroeder (2007), and Way (2009). Advocates of De Re Views will often use conditionals to capture their claims. For example, Schroeder (2007),14 holds: "For R to be a subjective reason for X to do A is for X to believe R, and for it to be the case that R is the kind of thing, if true, to be an objective reason for X to do A." But these conditions hold for Adina (trivially) and yet she does not have a reason (this holds for counterfactual conditional formulations as well). I discuss these problems some more in section 6 below.

A view that requires such a condition is a *de dicto view* of having reasons, because according to this view, in order to have a reason, you need to represent there being a reason *under this description*, under the description of it being a reason. [114]

De dicto views always have the same problem. Barring complex views about concept possession that we shouldn't saddle our account with, they over-intellectualise.[115] Young children, some animals, and people whose concepts don't include that of a "reason" plausibly have reasons and react on the basis of them. But they don't always (or never) have de dicto representations of reasons. One character from the literature serves as an especially strong example of this: Huck Finn. [116] Recall that Huck Finn fails to turn his friend in to the slave traders despite his conscious belief that he should, because slaves are property. Huck is perhaps somewhat bewildered by his own action, but it is, it seems to me, not completely beyond intelligibility for him. Huck reacts on the basis of a reason, so somehow Huck's action appears favoured to him. It is just very implausible that Huck believes he has a reason de dicto not turn Jim in. That is, it is not very plausible to represent Huck as thinking: 'Well, the fact that my friend is a human who deserves his freedom is a really good reason not to turn him in'.

The story of Huck Finn points away from de dicto views. This is because Huck Finn, in a sense, is the reverse of Adina, who fails to latch onto the normative significance of the fact that the hotel is on fire. Even though the fact that the hotel is on fire is a reason to leave the hotel and, let us suppose, Adina has the concept of a reason in her repertoire, it would be weird if Adina treated the fact that the hotel is on fire as a reason to leave it. That is, it would be weird for her to react in the way that the reason provided by the fact that the hotel is on fire would recommend, i.e. in her epistemic position, it would be weird for her to leave the hotel because it is on fire. Given her perfectly legitimate information, the relevant fact does not appear like a reason to her – and this seems sufficient to make her unresponsive if she nevertheless treated it like a reason to leave. Huck Finn on the other hand *does* treat the relevant fact like a reason – that is, he reacts in the way it recommends while being guided by it – even though he does *not* represent it as a reason and indeed cannot represent it as such given his background racist beliefs.

---

[114] Scanlon (1998), 25 holds such a view, according to which it is irrational to fail to respond to what one 'judges' to be a reason. Kolodny (2005) makes a similar assumption. Sylvan (2015), 590 summarises the view as follows: "De dicto: R is an apparent normative reason for S to φ iff it appears to S that R is an objective normative reason to φ."

[115] Parfit (2011), 118 says: "We can have rational beliefs and desires, and act rationally, without having any beliefs about reasons. Young children respond rationally to certain reasons or apparent reasons, though they do not yet have the concept of a reason. Dogs, cats, and some other animals respond to some kinds of reason...though they will never have the concept of a reason. And some rational adults seem to lack this concept […]"

[116] Apraly (2000); Audi (1990)

The correct view of the additional condition needed to "have a reason" must then lie somewhere between Adina and Huck in terms of the way in which reasons are represented in having reasons. It is this thought that directs us towards what might be called "guise views", according to which, roughly, the missing condition is that the relevant fact is presented to the agent under a "normative guise". A normative guise is more than merely representing the relevant reason de re, as in Adina's case. But it is less than having a representation with reasons as part of its content (see Gregory 2013).

Representations of propositions or facts that can count as reasons the agent has will be propositions that make their corresponding actions appear as if something was speaking in favour of them. They are representations that represent their relevant propositions under a certain light and thus fall under the class of seemings. When it seems to me that p, then p is presented to me as if p was true. When it seems to me that p is a reason to φ, then p is presented to me as if p spoke in favour of φ-ing. Just like when the stick in water seems to me as if it was bent, I do not have to have the corresponding belief with the content that the stick is bent, when p seems like a reason to φ, this does not require the belief with corresponding content that p is a reason to φ. Instead, the seeming represents the factual content of the relevant presentational state as speaking in favour of the corresponding action.

To borrow a phrase from an account similar to mine, the agent under such conditions will be *rationally attracted* (Sylvan 2015, 602) to treating p as a reason to φ. It will seem to them that p is good grounds for φ-ing. With my account of reacting for reasons already on the table, we can also say what it is to treat some p as a reason. To treat p as a reason is to be disposed to react on the basis of p.[117]

So, now we have clarified the representational elements of what it is to have a recognitional stance towards reasons. To have a recognitional stance towards p as a reason is to represent p under the guise of reasons (to be rationally attracted to treating p as a reason).

But this isn't enough. This is because the guise of reasons comes cheaply. Severely irrational agents like agents steeped deeply in "conspiracy think" and racists with a paranoid worldview in the background may (i) represent certain propositions and (ii) be rationally attracted to treating them as reasons. Sure, often completely irrational agents don't even treat the propositions that are presented to them as reasons, because they act on compulsions (they compulsively believe someone is after them for no reason at all). But I find it plausible that there are severely unresponsive agents who are more cognitively incompetent than practically incompetent. Conspiracy theorists can

[117] Technically, I am still talking about *sufficient* reasons here.

sometimes be recognizably deluded about what reality is like, but still have strong wills. These agents will react on the basis of propositions that seem to them like reasons, and consequently will be rationally attracted to treating them as reasons. But we should not say that these agents respond to reasons – so they should not count as taking a recognitional stance towards reasons.

This is why Recognitional Stance includes the condition that the representation under the guise of reasons itself must be the manifestation of a capacity. The idea is that propositions represented in responding to reasons must be represented competently, that is, the representation must be the manifestation of the capacity to respond to reasons. The agent being rationally attracted to treating the consideration that p as a reason must itself be a manifestation of the capacity to respond to reasons in order to count as part of the agent's recognitional stance. Recognitional stances are seemings competently brought about, as it were.

This condition excludes exactly those agents that are beyond the pale speaking in terms of responsiveness: the severely delusional and paranoid, whose actions/attitudes are not even bad responses to reasons. These agents plausibly do not exercise their capacity to respond to reasons at all in representing certain propositions under the guise of reasons.

It is important here to consider the larger picture of what this means philosophically. De Re views on having reasons are implausible because they completely ignore the agent's perspective in spelling out the epistemic side of responsiveness to reasons. This is implausible because, as Adina's case shows, the agent's perspective plays an important role in which agent can plausibly count as having reasons. Adina does not have reasons because she has no first-personal access to the reason there is for her to leave the hotel and, moreover, has first-personal assurance that the fact that the hotel is on fire isn't a reason. De Dicto views are implausible because they overemphasize the agent's perspective by building it into the very content of the relevant representational mental states. As the case of Huck Finn shows, the agent may respond to a reason even when they don't have a state with such a content and indeed even when they cannot have it. Guise views try to navigate this straight by locating the way reasons are represented in recognitional stances in an *appearance* of reasons. This property, they assume, marks proper responses to reasons. The thought is that when we respond to reasons, there is something it is like to respond – this is the guise of normativity, or guise of reasons. But as the cases of severe unresponsiveness show, this can't be right either. For the relevant appearance alone does not distinguish between these severe unresponsive cases and cases of legitimately picking up reasons. In both cases, a kind of appearance of reasons is present.

Thus, finally, I propose to narrow down the cases by introducing the condition that the appearance itself must be part of the manifestation of the right capacity in order to do the work normal guise views require it to do.

5.3. The Pieces Put Together

The point of the preceding discussion was to show that my exercise account of responding to reasons can helpfully unify a number of key cases in the vicinity of reasons-responsiveness. Put together, what I am proposing then is an account of responding to reasons which involves the crucial notion of an exercise (or manifestation) in at least three nodes in the process translating reasons into action/attitude (Mantel 2017, 2018 calls such an account 'triply dispositional'). Schematically:

Responding Schema: R => b(R) => M(R) => A

Where "=>" stands for the exercise of the RR capacity, "R" is the proposition that provides the relevant reason, "b(R)" is whatever mental state or stance instantiates the recognition of this reason, "M(R)" is the motivation to act for that reason and "A" is the action/attitude.

To respond to a reason then is to take a recognitional stance towards a proposition which is represented competently under the guise of reasons and to react guided by a proposition so recognised such that the reaction counts as an exercise of the capacity to respond to reasons.

As you can see, I have addressed mainly questions about M (R) => A and questions about R => b(R) here. Let me therefore say what I have chosen to omit from the discussion and why I find these omissions acceptable:

First, I have talked only about the capacity to respond to *sufficient* or *decisive* reasons, remaining largely silent about how the judgment that an agent has a sufficient reason to φ is pieced together from judgments about having some (but not necessarily sufficient) reasons to φ. Here is a sketch of how I think a full account will have to incorporate weights of reasons: The capacity to respond to reasons will have to include the capacity to respond to the relative weights of those reasons, that is, a capacity to recognize how weighty the reasons given in a situation are in relation to the other reasons present. I take it that sometimes no reason on its own is sufficient, so the agent will need the normative competence to correctly assess reasons-strengths. Other times, the reasons picked up will themselves already be sufficient, in which case their weight will have to be recognized without a weighing procedure. But the intricacies of the weights of reasons would take us too far afield here, so I will continue to talk about cases where there is one

clear sufficient reason the agent is able to recognize and translate into action via their capacity.[118]

Second, I have not addressed questions about motivation, that is, questions about what type of states must be involved in M(R). The controversy over M(R) is whether we need so-called conative states, typically called desires, in order to be motivated by reasons we have recognized. Let me again briefly hint at what I think my stance would be towards this discussion.

At the centre of at least the traditional debate is the discussion about whether it is possible to be motivated by beliefs alone. Defenders of this thesis point out that special types of belief, normative beliefs, for example the belief that I ought to pick up my son from school do "by themselves" motivate in virtue of their content.[119] Others reply that only conative states are able to precipitate action (Smith 1994, most famously), sometimes arguing that normative beliefs, in virtue of their motivating power, just are desires wrongly identified as beliefs.[120]

I think the view of responding to reasons based on exercising capacities I am here developing has the potential to helpfully circumvent the discussion about desires vs. beliefs in motivation, because what the exercise view insists on is that as long as the underlying realized phenomenon is that of exercising the capacity to respond to reasons, the realizer of this phenomenon is not important. I see no reason not to think that both desires and "normative beliefs" can count as exercises of the capacity to respond to reasons, so really the discussion about whether desires or special beliefs can motivate is moot. It is possible that both do. Moreover, in terms of the philosophy of mind involved in the debate, it might even be helpful to pick out those states that do push towards action in the right way simply as exercises of capacities, rather than "desires" or "beliefs" thereby dissolving a debate that already has the distinct air of being verbal.

Of course, what I also haven't discussed is what it *is* to exercise a capacity. But as I already pointed out in some places, chapter 7 will be dedicated to this.

While the account should look attractive in its own right, part of its appeal is also that it rectifies the mistakes other approaches to responsiveness make. Most interestingly

[118] For an overview and the development of some issues and views on weighing reasons, see Lord & Maguire (2016) and Cullity (2019).

[119] See for example Nagel (1970), McDowell (1979), Platts (1980), McNaughton (1988), Dancy (1993), Scanlon (1998), and Shafer-Landau (2003).

[120] Other, more complicated views exist, including whatever exactly Wedgewood (2004) advocates for and the converse view that all desires are just beliefs about reasons (Gregory 2017) and the notoriously elusive suggestion that motivating states are 'besires' – states with both directions of fit (the label comes from Altham (1986), but advocates include McDowell 1995 and Wiggins 1987).

perhaps, it rectifies the mistakes made by both other Univocal Views and Non-Univocal views, striking a balance that will hopefully be found appealing by many except the staunchest defenders of Non-Univocal Views. To show this, the next sections, which comprise the second part of this chapter, discuss problems for Univocal and Non-Univocal Views.

Part II
Problems for Other Proposals

6. Problems for Univocal Views

Recall that central to the idea behind Univocal Views is the factoring picture, according to which we can locate a basic relation in Achievement, Fact-Guidance, and Consideration-Guidance, from which the others can be derived. As we saw, if we think that Fact-Guidance and Consideration-Guidance involve the same relation – a kind of intelligibility explanation in terms of considerations – then there will be only two ways to conceptualise Univocal Views: (i) the underlying relation is explanation in terms of believed-to-be-true considerations (which Lord refers to as the relation we stand in when we react for 'motivating reasons') (ii) the underlying relation is that of reacting for normative reasons.

Now, since my account is already on the table, we can see that we *should not* assume that Fact-Guidance and Consideration-Guidance cluster together in terms of the relevant relations. Nor should we assume that (i) and (ii) are the only Univocal options possible. After all, my account shows that the plausible third option is that *neither* is primary. Instead, the unifying element between Achievement and Consideration-Guidance is the exercise of the capacity to respond to reasons (which is not present in Fact-Guidance). Nevertheless, my view will be strengthened by seeing how exactly the other two options fail.

(i) Consideration- Guidance First

We have already encountered what I take to be the most common Univocal View, which I will call, Consideration-Guidance First (Lord calls it Motivating First). Consideration-Guidance First views claim that the relation that underlies both Achievement scenarios and Fact-Guidance and Consideration-Guidance scenarios is that of reacting for believed-to-be-true Considerations.

The idea is best illustrated by focussing on Rita in the Achievement scenario and Boris in the Consideration-Guidance scenario, in which Boris is faced with a deceptive environment. Even if there is a special achievement explanation available for Rita, it seems whatever explanation is available for Boris will also be available for her. In particular, both react on the basis of believed-to-be-true considerations, it seems. Reacting for normative reasons then just is reacting for believed-to-be-true considerations– considerations that are also true and normative reasons.

We have already seen the argument for why this position fails. In cases of accidentality like that of the Fortunate Consequent Affirmer, an agent reacts for a believed-to-be-true proposition (a proposition that is in fact true) and this consideration happens to be a normative reason for reacting in that way. But we still wouldn't say that the agent responds to the relevant reason. Sam still believes what he has reason to believe by accident.

Again, this failure can be expressed as a failure of the factoring picture: What the argument shows is that the relation in Achievement isn't just the relation common Fact-Guidance and Consideration-Guidance plus a success condition. This is because the Fact-Guidance scenarios exist. In those scenarios, the agent reacts for believed-to-be-true considerations and the success condition is flicked on. But we don't get the same scenario as in Achievement. The Consideration-Guidance First idea relies on the reacting for normative reasons is reacting for believed-to-be-true considerations + success picture. So Consideration-Guidance First fails. It fails, that is, unless we realise that the relation which Consideration-Guidance has in common with Fact-Guidance is not a relation relevant to responsive agency, because we get basic intelligibility even from crazy agents. The important relation in Consideration-Guidance is what it doesn't have in common with Fact-Guidance: the exercise of the RR capacity.

(ii) Normative-First

Normative-first approaches turn the direction of explanation around. They assume that the more fundamental notion is that of reacting for normative reasons. The idea is roughly that reacting on the basis of believed-to-be-true considerations is a weak or impoverished form of reacting for full blown objective normative reasons. An attitude like this may be detected in Davidson's famous (and infamously puzzling) quote that motivation is a kind of "anaemic justification" (Davidson 1963, 691) and in the often repeated dictum that motivating reasons provide rationalizing explanations. I take it that the traditions of conceiving action as done "under the guise of the good"[121] also falls under this category, even if this view is often characterised as one about the nature of action more generally.

The view takes several different shapes. Arpaly & Schroeder (2014) base their theory of responsiveness on what they call "rationalizing reasons", which are contents of beliefs tied to intrinsic desires, which although false, still play the crucial rationalizing role in

[121] Prominent characterisations of acting under the guise of the normative as the guise of the good are found in Anscombe (1963), 75 and Davidson (2001), 22-3, as well as Velleman (1992b)

explaining action (see Arpaly & Schroeder 2014, especially sections 3.1 and 3.5). The view passes as a Normative-first Univocal View because Arpaly and Schroeder are quite clear that the rationalizing relation is the common element in reacting for reasons and that rationalisation makes agents at least to some extent rational – a power prima facie reserved for real normative considerations. But there are several aspects of their view which would lead this discussion too far afield, chief among them the fact that the view is a desire-based view of responsiveness, which, to name just one problem, is not generalizable to believing for reasons (because it seems quite implausible that we believe on the basis of intrinsic desires). For this reason, I leave the Arpaly & Schroeder view to one side. I direct you to Lord (2018), ch. 6.4. for a treatment that fits into the current dialects, at least to some degree. Another view that seems to take reacting for normative reasons as basic is Ralph Wedgwood's account of reasoning (Wedgwood 2006). But I will not engage with the account here.[122]

I shall focus here instead on the most clear and well-developed account of the Normative-first View. This type of view holds that the core notion common to both reacting for normative reasons and acting for believed-to-be-true considerations is that of acting for a 'subjective' or 'apparent reason'. Apparent reasons are considerations that are (i) presented to the agent as true and (ii) would be objective normative reasons if they were true. (Parfit 2011, Schroeder 2007, Way 2009)

There is a lot of internal struggle over both conditions however, especially over the question whether we should put the second condition in terms of a subjunctive conditional (see Schroeder 2007, Schroeder 2009, Whiting 2014, Sylvan 2015). I will run with this quite neutral and barebones definition until the question becomes relevant.

The apparent reasons idea need not subscribe to the factoring picture that brought down the Consideration-Guidance First view. This is because the idea can be described as holding that there are two distinct reasons-relations an agent can stand in. Agents like Rita in Achievement are related to real objective reasons. Agents like Boris in No Mistake Consideration-Guidance are related to apparent reasons. Both count as responsive to reasons, notably, because both real normative and apparent reasons are reasons proper. There is of course a remaining residue here from the factoring picture in that apparent reasons are appearances of real reasons in a minimal sense: in the sense that they *would be* objective normative reasons if they *were* true.

The apparent reasons solution makes sense if we think about the distance between Rita in Achievement and Boris in No Mistake Consideration-Guidance (recall that Boris and

[122] Wedgwood has no explicit account of how reacting for normative reasons will give you reacting for motivating reasons, he only refers to Grice (2001) in a footnote (the Grice account seems compatible with my account of exercising capacities badly).

Rita are intrinsic duplicates, but Rita's hotel *is* on fire and Boris's *isn't*). This is because the way in which Boris is almost like Rita, and negligibly unlike her, can very plausibly be described by saying that everything appears to Boris as if there is an objective reason for him to believe the hotel is on fire and leave the hotel (which also prima facie supports the subjunctive conditional in the second condition above: if the proposition that the hotel is on fire was true, and it very nearly is, it would be a good reason to believe the hotel is on fire and to leave it).

Non-Univocal Views about responding to reasons are achievement views. They focus on the fact that Rita from Achievement gets things right in virtue of her capacity to get them right. Apparent reasons accounts, unlike Consideration-Guidance First accounts, are able to retain some of the elements that make achievement views attractive. Responsiveness, they claim, is a matter of possible achievement, or more precisely, achievement in the closest possible world where the conditions are right.

However, the apparent reasons view has a lot of trouble getting the distances between our other sample agents right. Take Hannah from Tiny Mistake Error Case. Hannah has made some mistake in recognizing the important facts (she recognized the fact that she was told about the fire drill but did not keep it in mind). So she should count as less responsive than Boris. But Rita also has apparent reasons – things are presented to her as if the hotel was on fire and if things were presented to her veridically, there would be an objective reason for her to leave. We might think that the worlds in which Hannah's reactions are an achievement are not as close as the worlds in which Boris's reactions are an achievement, but the precise closeness-measure that would render this result is not entirely clear. The demon might be present, and in the mood for shenanigans, in all of the worlds where Boris is in this kind of situation. So modally, Boris's achievement might be quite far away. However, epistemically Boris is still closer to Rita than Hannah is. Modal distance just isn't the right measure to get the differences between Hannah and Boris right.

But even if the right measure of closeness could somehow be fixed, a larger problem would remain. This is that the apparent reasons solution also has problems with Sam, the Fortunate Consequent Affirmer from the Fact-Guidance case. When Sam believes guided by the fact that Terry's car is in the driveway, he still doesn't respond for the reason that Terry's car is in the driveway. But the apparent reasons account predicts that he does. After all, the proposition which is a reason for believing Terry took the bus to work is presented to him and if it was true it would be an objective reason if it was true. The proposition *is* true, after all, and so it *is* an objective reason. The subjunctive conditional is trivially true. So Sam counts as responding to the (apparent) reason that Terry's car is in the driveway, which is precisely what we don't want to say. The fact that

Sam is completely off when it comes to the normative bearing of the consideration he latches onto is not mitigated by the fact that the consideration would be a reason if it was true. For even if it *is* true, Sam does not recognise its normative significance. The core problem can once again be described as one of accidentality. Sam latches onto a fact, and it is an accident that this fact is also the reason for believing as he does. The same is true for the counterfactual condition. Agents may latch onto considerations or facts of which it is true that they would be reasons if they were true but do so accidentally. A notion of apparent reasons which allows for this possibility will be neither here nor there in terms of responsiveness to reasons, for in order to respond to reasons, agents have to *competently* (non-accidentally) latch not their normative bearings.

What is needed, apparently, is what I called a 'competently formed guises' view of reacting for reasons (See Sylvan 2015 and my own account above), according to which the appearance of reasons is itself the manifestation of the capacity to respond to reasons. Sam fails this condition.

For these reasons the standard apparent reasons solution, the second direction of the Univocal View, fails. It would seem as if both ways to defend a Univocal View have failed. And so both traditional ways of developing a Univocal View fail. Both directions of finding a common core of responding to reasons fail moreover, because they count agents as responsive who are not responding to reasons.

Univocal Views are permissive by nature, because they assume a great variety of cases can be subsumed under one philosophically informative heading. But it seems that the Univocal Views we have encountered so far are too permissive, with no way of restricting themselves without losing their defining aspect: the assumption that agents like Rita, Boris and Hannah have something important in common. I have shown above that my account can hold on to this insight (what they have in common is an exercise-explanation in terms of the capacity to respond to reasons).

Unsurprisingly, Non-Univocal Views have the opposite problem. They are too restrictive. I discuss these problems in the next section.

7. Problems for Non-Univocal Views

Non-Univocal Views claim that responsiveness is drastically split. Achievement, Fact-Guidance, and Consideration-Guidance only seem to have common elements, according to these views, either because philosophers have latched onto superficial commonalities or because they have not seen clearly enough that genuine achievement explanations are distinct, (or both). As I see it, there are at least two types of Non-

Univocal View, the difference being dependent on where they think the split occurs in responsive agency.

What I have called Achievement Views (Lord 2018, probably Kiesewetter 2017) assume that the split occurs between cases of genuine normative achievement and cases of lack of achievement. Since agents may not enjoy such an achievement and yet act guided by facts (Sam the Fortunate Consequent Affirmer), the view singles out explanations by facts through capacities as distinctive.

Another type of view, which might be called Fact Guidance View (Alvarez 2010, 2018, Cunningham 2019a, Littlejohn 2012, ignoring some subtleties) assumes the split occurs between responses guided by facts and responses guided by false propositions.[123] Fact Guidance Views result from restricting even the terminology of 'motivating reasons' to a factive use, in the sense that if S acts for the motivating reason p, then p. Since they take this position to be crucial, they single out guidance by facts as distinctive. This is often paired with the harsh assumption that agents who act on the basis of false propositions act "for no reason at all" (Alvarez 2018, Littlejohn 2012). The view is also often put by claiming that false propositions are "counterfeit reasons" (Littlejohn 2012, 127) – they stand in the same relation to objective normative reasons as rubber ducks do to ducks.

It is important here to pinpoint the specific difference between the counterfeit reasons of the Fact Guidance View and what I discussed earlier under the label of 'apparent' or 'subjective reasons'.[124] The two accounts are importantly different (they belong two opposing camps about whether responding is Univocal, for one), but are often not clearly distinguished. Alvarez even uses the term "apparent reason" for her rubber duck reasons, crediting Parfit in a footnote. Yet it is quite possible that Parfit had a substantially different idea in mind when he talked about apparent reasons. What is the difference? The difference is that apparent reasons as the term is used in Normative-first Univocal approaches are in fact, in some sense, reasons. That is, they genuinely favour reactions – making them morally right, justified or rational. This is especially explicit in Schroeder

[123] This disjunctive conception leads Alvarez to a disjunctive conception of rationality according to which an agent is rational if they correctly respond to either real reasons or apparent reasons (where these two relate to each other in a rubber duck way). The issues of Non-Univocal show up here again. Why, we should ask, is an agent's rationality grounded in both real and apparent reasons, given that they are not both real favourers?

[124] The terminology is complicated further by positions most prominent about moral obligations according to which what agents have an moral obligation to do depends in part on their epistemic perspective, for example Jackson (1991); Prichard (1932); Ross (1939, 146–67); Scanlon (2008, 47–52). See Kiesewetter (2018) for a discussion.

(2008), who argues for what he calls two "reasons relations"[125] and Comesaña & McGrath (2014), who at one point claim:

> On our view, reasons [one has] are considerations—true or false—that favour or support a person doing (believing, feeling, etc.) something. (2014, 76–77)

Counterfeit reasons cannot do this. By definition, they are faking normativity. This is why Fact Guidance Views are Non-Univocal while apparent reasons accounts are Univocal – because for the apparent reasons account, both reacting for apparent and reacting for objective reasons is reacting for a type of reason, while for the Fact Guidance View, only the latter is.[126]

Fact Guidance Views and Achievement Views are not just notational variants of one another. They differ in how they treat Sam, most importantly. The Achievement View treats Sam as removed to a significant extent from responsive agency because his reactions accidentally align with his reasons. The fact that he has explanations available in terms of facts does not change this. The Fact Guidance View on the other hand would insist that Sam is still closer to responsive agency than agents who act on the basis of false propositions – for such agents act for no reasons at all.

It is plausible though that Fact Guidance views can take the lessons of Achievement Views on board, so that they can hold the position that responding is split into three distinct relations: genuine achievement, guidance by facts, and explanation by no reason whatsoever.

All types of view have problems with our cases however. The problems they have are so closely related that I will discuss both Achievement and Fact Guidance Views in the same section, pointing out crucial differences where it is necessary.

To see where Non-Univocal Views go wrong, take Sam the Fortunate Consequent Affirmer again. The important truth about Sam is that he believes what he has reason to believe by accident. Following his own strange system of rules, he stumbles upon an instance in which these rules happen to overlap with what normative reality recommends. This makes Sam very unresponsive, much more unresponsive than agents like Hannah (Tiny Mistake Consideration-Guidance) and Boris (No Mistake Consideration-Guidance), who are exercising their capacities but fail in some respect.

---

[125] This also shows why Schroeder can reject the factoring picture without giving up dedication to a Univocal View. He thinks having reason is importantly unified in that both having "subjective reasons" (false propositions) and having "objective reasons" (true propositions) is standing in a reasons-relation.

[126] The clash becomes especially obvious in Alvarez (2018), 3350, in which the Alvarezian terminology of apparent reasons ("My claim is that a false, or in my preferred terminology, an apparent motivating reason is not merely a bad reason but no reason at all") is contrasted with Hornsby (2008) and Turri (2009), who both seem to hold that apparent reasons are, in some sense, still reasons.

So Sam should count as less responsive to the reason in his situation than Hannah, who is diligent but makes a tiny mistake. Being right accidentally is worse than being wrong (in the sense in which this is attributable to Hannah), in terms of assessing the attunement of an agent to their normative situation. If this is not immediately obvious to you, notice that two indicators for how in tune an agent is with the normative facts clearly point towards it. First, if you believe that rationality is a mark of responsiveness to reasons, then you should believe Sam is worse off than Hannah. For Hannah surely is, even if far from fully rational, far less irrational than Sam. Second, if you think that closeness of possible achievement is a measure of responsiveness (as I alluded to above), then you should think that Sam is worse off than Hannah. This is because we need to imagine Sam without his weird and possibly deeply entrenched dispositions in order to imagine him as genuinely responding to reasons. We only need to imagine Hannah a little more alert or a little more careful with her beliefs. The idea that Hannah is more responsive than Sam should in general be easy to swallow. No matter what view of reasons, responsiveness and rationality we have, we should be able to agree that accidentally responding is not responding at all.

Unfortunately, the Fact Guidance View does not vindicate this judgement. This is because while Hannah is not guided by a fact and therefore acts for no reason at all, Sam is guided by a fact and therefore has an explanation in terms of at least motivating reasons available. This is exactly the opposite of the intuitive judgment. Now Hannah is on the outer fringes of what we can still reasonably explain while Sam is much more in the centre. This seems like a severe flaw of any theory about responding to reasons. If there is a categorical gap anywhere in the spectrum of agents I am working with here, it should be between Rita, Boris and Hannah on the one side and Sam (and the ever more crazy versions of him) on the other. When our theory renders the result that an agent who is completely blind to the normative features of their situation counts as closer to full responsiveness than an agent who tries their hardest and almost succeeds, we should be deeply suspicious of the structure of this theory.

The structural aspect of the Fact Guidance View which causes this result is its adherence to the position that acting on the basis of false propositions is not acting for reasons at all, not even motivating reasons. Here is Clayton Littlejohn professing such a view about an agent who runs away because they falsely believe a serial killer is chasing them:

> I agree that we can explain Leo's action by saying that he runs because he believes falsely that he is being chased. I agree that there are reasons why he acted as he did. I agree that when we see what these reasons are, we can see why Leo was perfectly reasonable in acting the way he did. What I deny is that that Leo acted for a reason. There was nothing in light of which he did what he

> did. We all know why Leo ran – he ran down the hall because he believed that the killer was after him. This does not explain his action in terms of motivating reasons because it does not tell us what his reasons were – it turned out that he had none. The reasons why Leo acted as he did are facts about Leo's mental states and these provide us with a perfectly good explanation of Leo's behavior. We regard Leo's actions as reasonable because we accept a causal explanation of his behavior according to which Leo's behavior was controlled by the mechanisms responsible for responding to reasons and because we think he was not unreasonable in taking himself to have the reasons that would speak in favor of running. (Littlejohn 2012,155)

The quote is puzzling for a number of reasons, including the curious remark that Leo's behaviour was controlled by the mechanisms of responding to reasons, which is more than a bit surprising if he literally acts for no reason at all. Most importantly however, it just seems plain false that there is nothing in light of which Leo acts. Listening to Littlejohn, it sounds like the only explanation available for Leo's behaviour is one in terms of explanatory reasons. That is, we can explain his behaviour third-personally with recourse to his belief that a killer is chasing him. But what about Leo's perspective? His belief becomes transparent first-personally and it seems we can easily pinpoint the proposition in the light of which he acts: it is that a serial killer is chasing him. It is because this proposition guides Leo's behaviour through the exercise of his capacity to respond to reasons that we can assess his behaviour as reasonable.

If we did not have this explanation available, then Leo would be identical in terms of how we explain his actions to characters like Radioman, who compulsively turns on radios without being able to say why (Quinn 1993). Radioman acts for no reason at all because there is nothing in the light of which he reacts. Further, we can understand Radioman's behaviour via a causal explanation in terms of Radioman's mental states – his disposition to turn on radios. But Leo is not like Radioman. And by extension, Hannah (and for that matter Boris) is not like Radioman. There is something that makes Hannah's reaction perfectly intelligible to herself from her perspective: the proposition that the hotel is on fire.

The Achievement View has the resources to calm these worries. This is because the Achievement View can point out that Sam is as far away from an achievement as an agent can possibly be. After all, Sam's success is accidental, and this is exactly what an achievement does *not* consist in. Hannah's failure on the other hand is at least *her* failure. So, measured in terms of closeness to an achievement, Hannah is much closer to Rita than Sam is. Additionally, an Achievement View is not committed to the thesis that acting for false propositions is acting for no reason at all. In fact, Lord's Achievement View

directly denies this (Lord 2018, 152). For him, reactions for motivating reasons might be reactions on the basis of false propositions.

But the Achievement View has another problem. The problem is the distance between Boris from the No Mistake variant of the Consideration-Guidance scenario and Rita from Achievement. Clearly, Boris is very close to Rita in terms of his attunement to his normative situation. Rita and Boris are intrinsically alike. It is very plausible that we would evaluate them as rationally on a par. Further, as stipulated, Boris fails to react for normative reasons through no fault of his own, so it seems Boris even did everything Rita did (again, they are intrinsically alike) in terms of at least trying to respond to reasons.

So, in terms of accounting for the actual phenomena we are trying to capture, it would be regrettable if we had to sort Rita and Boris into two different und fundamentally distinct categories. Yet this is exactly what the Achievement View does. It says that only Rita has an achievement explanation available to her. An account of responsive agency that cannot find a deep similarity between Boris and Rita will on balance be inferior to an account which manages to find such a similarity.

In order to reinforce the impression that there must be something connecting Rita and Boris, think about their relative distance to Hannah. If the Achievement View is right, then Hannah and Boris are at the same level in terms of their responsiveness. Both fail to respond to reasons. So both are at an equal distance to Rita, who does respond to reasons. But the intuitive judgment about Rita, Boris and Hannah is that Boris is closer to Rita than he is to Hannah, or in other words, that the distance between Rita and Hannah is bigger than the distance between Boris and Rita. Even more problematically, imagine several iterations of Hannah, each making a more severe mistake. This will gradually liken Hannah to Sam, taking her more and more unresponsive versions further and further away from Boris and Rita. But the Achievement View will only render the judgment that we are increasing the distance between Hannah and Rita. For it will hold that in terms of the crucial achievement relation, all iterations of Hannah are just like Boris: unsuccessful. This makes it look like on the Achievement View Boris (who is equidistant to all versions of Hannah in terms of achievement) is very much like Sam. But it seems to me like this is the opposite of what we want. Sam is thoroughly unresponsive. He is not in tune with his normative situation. Boris is very much in tune. It is just that his situation sends him deceptive signals.

The advocate of the Achievement View might argue that in making Hannah more and more like Sam, we are increasing the modal distance between achievement and the agent's current situation. But the intuitive judgment is that we are thereby also increasing the distance between Hannah and Boris. However, the modal strategy clearly does not

render this judgment. For modally, Hannah might be much closer to Rita than Boris is to Rita. That is, the worlds in which Hannah achieves success in responding are much closer to her actual situation than the worlds in which Boris achieves success are to his current situation. So consequently, the distance between Rita and the ever more severely unresponsive versions of Hannah will not be proportional to the distance between Boris and these versions of Hannah. However, we *should* expect versions of Hannah in which she makes more and more severe mistakes to be proportionally as far away from Boris as they are from Rita. After all, Hannah is intuitively becoming more unresponsive, while Boris and Rita stay the same.

We can now see how Non-Univocal Views face the opposite problem of Univocal Views. Univocal Views are too permissive. But Non-Univocal Views are too strict. They obliterate the commonalities that there clearly are between successful and unsuccessful agents, thereby making a unified conception of responsive agency impossible.

My Exercise Univocal account fares better in this respect. It gives success and error their rightful place, thereby occupying the correct middle ground between Univocal Views and Non-Univocal Views and getting the distances between Rita, Boris, Hannah, and Sam right.

Let me finally come back to a point I hinted at earlier: that we might take the attitude towards the preceding discussion that it is merely verbal.

8. Have We All Been Talking Past Each Other?

The philosophical literature on responding to reasons is filled with long debates about whether specific notions like "possessing" or "having" and "acting for" reasons are factive. But it contains very little direct engagement with what seems to me clearly the underlying issue – namely the question of how unified our account of responsive agency ought to be. I have therefore tried to address this issue directly in this chapter. But while assessing the debates about factivity, you might sometimes get the feeling that the disagreement is less drastic than it might seem, precisely because those who insist that "acting for reasons" is factive often seem interested in a higher level phenomenon while those who insist that it isn't factive seem more interested in a lower level phenomenon. In terms of my account, factualists are often more interested in specific types of exercises of the capacity to respond to reasons – factive exercises –, while non-factualists often seem interested in the more general phenomenon of responding, perhaps.

Just take Littlejohn's puzzling quote from the previous section again in which he describes what is going on when an agent flees on the basis of their false belief that a serial killer is chasing them:

> I agree that we can explain Leo's action by saying that he runs because he believes falsely that he is being chased. I agree that there are reasons why he acted as he did. I agree that when we see what these reasons are, we can see why Leo was perfectly reasonable in acting the way he did. What I deny is that that Leo acted for a reason. There was nothing in light of which he did what he did. We all know why Leo ran – he ran down the hall because he believed that the killer was after him. This does not explain his action in terms of motivating reasons because it does not tell us what his reasons were – it turned out that he had none. The reasons why Leo acted as he did are facts about Leo's mental states and these provide us with a perfectly good explanation of Leo's behavior. We regard Leo's actions as reasonable because we accept a causal explanation of his behavior according to which Leo's behavior was controlled by the mechanisms responsible for responding to reasons and because we think he was not unreasonable in taking himself to have the reasons that would speak in favor of running. (Littlejohn 2012, 155)

As I said, it seems a bit surprising that Leo turns out be literally acting for no reason on Littlejohn's account even though Leo's reasons-responsiveness capacity is active. Maybe Littlejohn isn't a full Non-Univocalist, as it were, because he might allow that there is some faint common underlying core to acting on the basis of true propositions and acting on the basis of false propositions.

Lord (2018. Ch. 6.6) offers a neat way of expressing this dialectical nuance. Non-Univocal Views are a form of disjunctivism. Lord points out that there are two ways of taking disjunctivist stance towards a certain phenomenon. Negative disjunctivism about responding to reasons is merely the negative thesis that what happens in Achievement involves a different relation from what happens in the bad cases. Positive disjunctivism about responding to reasons holds that there is a common phenomenon in both Achievement and Consideration-Guidance cases, but this phenomenon, responding to reasons, is itself disjunctive. If we hold positive disjunctivism, then we hold that reacting for normative reasons and reacting for motivating reasons each realize the disjunctive kind they belong to.

My account offered over the course of this (and the last) chapter can be interpreted as a kind of positive disjunctivism. After all, I am claiming that both reacting for normative reasons and reacting for believed-to-be-true considerations realize the phenomenon of

responding to reasons. Interpreted like this, my account is a peace offering to those arguing about the factivity of acting for reasons. Maybe the advocates of non-factivity have in mind a positively disjunctive phenomenon which may be realized in non-factive ways, while the advocates of factivity have already zoomed in on one type of manifestation of this phenomenon, which they correctly take as factive. This is in fact the route that Lord himself takes tentatively (See Lord 2018, p.178). So, was my stance Lord's view in this chapter merely for show?

I don't think it was. While I agree that responding to reasons is a disjunctive phenomenon in that both reactions based on false and based on true propositions count as manifestations of it, I deny that the phenomenon is essentially disjunctive. That is, the arguments I made here and the account offered rely on the assumption that exercising the capacity to respond to reasons, and the corresponding availability of an exercise-explanation of the relevant response always has the same explanatory core – the exercise relation.

I don't believe a Non-Univocal account can meaningfully accept this feature of what I argued for here. For what follows from this feature is a new vindication of the original Williams dictum that "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 1979, 102)

That is, it follows from the claim that exercises always involve the same explanatory and metaphysical set-up and the claim that both reacting for motivating reasons and reacting for normative reasons are exercises of the capacity to respond to reasons that the truth value of the propositions involved in an exercise-explanation is negligible. The core explanatory phenomenon then is not disjunctive, but unified. This, it seems to me, is a strong and undeniable Univocal kernel that my account retains and which allows it to explain the surface level phenomena in the way it does. Accepting my kind of disjunctivism therefore means accepting a Univocal account. Chapters 6 and 7 will elaborate on the explanatory features of exercising capacities that I will use to fully spell out the ideas developed here.

With this, the account of responding to reasons developed in this chapter is complete. It is an account of responding to reasons that (a) captures the important cases, especially cases of accidentality and it is (b) an account that does not outwardly rely on any modalist vocabulary. All we need to understand, according to the Exercise Univocal View, if we want to understand what it is to respond to a reason, is that exercise-explanations are available in the relevant cases.

However, since I have relied, in this chapter, on an intuitive understanding of exercising capacities and the closely connected availability of exercise-explanations, open questions remain. Most importantly, we might want an account of the explanatory structure – or 'mechanics', as it were – of what is going on explanatorily in exercise-explanations. We might want to know how they work. This question is especially pressing because I promised in chapter 3 to develop an account of non-accidentality in explanatory (as opposed to modalist) terms. All we know so far is that the availability of an exercise-explanation disperses the impression of accidentality – or in other words establishes a non-accidental relationship between the relevant items. But to know this is to know very little about the concept of non-accidentality.

The next two chapters are therefore dedicated to spelling out how exercise-explanations work. The next chapter will present a complete explanationist account of non-accidentality. Chapter 7 will reconnect this account to the notion of an exercise – i.e. it will show how the notion of an exercise is accidentality dispersing. These last two chapters can therefore be seen as spelling out the deep structure of the explanatory relations I have taken for granted in this chapter. In this sense, they complete the demodalising project developed in chapter three – the project of understanding reasons-responsiveness in explanatory terms instead of in terms of alternative possibilities.

Chapter 6:

An Explanationist Account of Non-Accidentality

1. Introduction

In the previous chapters, I presented an orthonomy view in which the notion of an exercise of the capacity to respond to reasons takes centre stage. One of the central jobs the notion of an exercise is performing in this account is that it ensures what I called in chapter 3 a non-accidental relationship between action and reason. But I have left it open just how exercising capacities ensures non-accidentality. This is because I treated the notion of an exercise as primitive in the previous chapters. This chapter and the next are designed to rectify these points. That is, I shall in this chapter present a theory of non-accidentality. In the next chapter, I will use this theory to shed light on the notion of an exercise.

The overarching aim of these twin chapters is to fulfil the promise I made in chapter 3: To develop a theory of non-accidentality that is explanationist, not modalist in nature, and to thus rid theories of reasons-responsiveness of their deep structural problems with Frankfurt-like cases. In order to develop such a theory, I will first have to say some things about the notion of a coincidence in general (sections 2 and 3). I will then move onto accidentality/non-accidentality (sections 4-7). Finally, the chapter also develops an argument against the modalist view of non-accidentality (section 8), an argument which has the additional advantage of systematically explaining how a host of accidentality cases are related. The argument ties together some of the themes of the thesis: I have already pointed out at several places that orthonomy notions can't seem to be decomposed into separate independent components. This principle will be at the heart of my account of non-accidentality. It is also why modalism inevitably fails – because it treats orthonomous/non-accidental relationships as ultimately decomposable into a plurality of modal truths. It is also why modalism is eternally faced with Frankfurt-like cases – because these cases exploit the possibility of the modal components of the modalist rendering of an orthonomous relationship holding independently of each other (which constitutes a kind of accident).

Let's refresh our memories. In chapter 3 I pointed out that there are two approaches to non-accidentality. Modalism holds that non-accidental relationships are roughly relationships of modal tracking such that something like this holds:

Modalist The fact that p is non-accidentally connected to the fact that q iff a sufficient proportion of the relevant p-worlds are q-worlds.

Explanationism holds that non-accidental relationships are roughly special explanatory relationships such that the following holds:

Explanationist The fact that p is non-accidentally connected to the fact that q iff q explains$_{unique}$ p.

Recall that unique explanations are those that are forthcoming if agents exercise their orthonomy abilities. When I know that p (instead of just truly believing that p), then we can uniquely explain my belief that p in terms of p (and my exercise of my relevant ability, but this is the topic of the next chapter). When I respond to a reason R, then my action can be uniquely explained in terms of R. It is important that these are 'unique' explanations because of the existence of cases in which R explains my action in *some sense*, but the relationship between my action and R still counts as accidental. These are cases of deviance, which I will discuss further below.

Uniqueness of course is a placeholder term right now. The project of developing an explanationist account of non-accidentality is the project of spelling out in exactly what sense the relevant explanations are unique. This chapter gives an account of uniqueness of the relevant explanations. Here is a brief anticipation of my eventual account: The relevant explanations are unique in that they cannot be decomposed into separate independent components. They are *unified*, as I shall say. To understand what this unification amounts to however, we have to start from the beginning.

2. Symmetrical Coincidence Questions and Coincidence

Accidentality is a special form of coincidence. Therefore, I will start by sketching how to think about coincidence as an explanatory phenomenon.

A good place to start are what might be called traditional[127] accounts of causal coincidence, which are summarised succinctly by David Owens: "[…] a coincidence is an

[127] Lando (2017) refers to these accounts as 'traditional'. Traditionalist tendencies can be found in Owens (1990), Owens (1992), Monod (1970), Horwich (1982), Lange (2010), and Sober (1984), although clear attributions are complicated by distinction between causation and causal explanation (discussed briefly in the next chapter).

event which can be divided into components separately produced by independent causal factors" (Owens 1992, p.13).

According to this sort of account, coincidences are composite events that have no cause (only their components do). For example, both me and you sneezing at the same time is coincidence iff there is no cause for this composite event, just a cause for your sneezing and a cause for my sneezing. However, your sneezing and my sneezing might have the same cause, for example the dispersion of sneeze particles in the air. In this case your and my sneezing cannot be divided into separate independent causal factors. Consequently, it will not seem like a coincidence that we both sneezed (at least not a complete coincidence, more on this below).

My account follows this core idea, but it abandons Owens strict focus on causal vocabulary. The deciding factor in the above example, it seems to me, is not that we gave a *causal* answer to the question whether my sneezing and your sneezing was a coincidence. The deciding factor is whether we can answer a type of why-question[128] in a specific way.[129]

Note that we can ask coincidence related why-questions about a wide variety of things, many of which do not fit the description of events.[130] For example: Why are we both wearing red shoes? Why are both pancakes shaped like Elvis Presley? Why are you both slowly blushing? To all of these questions we may add, in order to emphasise what we are asking: Is it a coincidence?

I shall understand the phenomenon of coincidence in terms of these why-questions, that is, in terms of what kind of answer can be given to them. These why-questions can be, but don't always have to be, answered by telling a causal story. They sometimes are, but don't always have to be, about two events. But they do form a distinct cluster. They are always about the relation between two facts which share some salient commonality. The question about the red shoes, for example, is about the relation of the fact that I am wearing red shoes to the fact that you are wearing red shoes, two facts that share the salient commonality of the red shoes. I shall refer to this type of question as *coincidence-questions*.

[128] Skow (2016) develops a theory of explanation in terms of answers to why-questions. This is significant because it shows that not only can we understand coincidences in terms of why questions, we can understand the notion of an explanation itself in terms of why-questions (and not in terms of for example counterfactuals).

[129] The impression that the traditionalist focus on causation should be replaced by explanatory vocabulary is shared by Lando (2017) and Bhogal (2020), who both advance explanationist accounts of coincidence.

[130] Further examples of coincidences we would not immediately describe as involving events from all over the philosophical literature, collected in Johnson King (2020), include Carr (1979) and Riggs (2014).

Coincidence Questions (Symmetrical): Why [p&q]? – where p and q share some salient commonality.

How can we understand coincidence in terms of the answers to this type of question?

It would be tempting to assume that *it is a coincidence that p&q iff there is no explanation as to why p&q*. But unfortunately, it is not quite so easy. For we can always explain why a conjunction holds by just conjoining the explanantia of its conjuncts.

For example, let us say that we are both watching Wim Wenders's masterpiece *Kings of the Road*[131] at the same time and place. Is it a coincidence that you are watching *Kings of the Road* and I am watching *Kings of the Road*? Well, we can simply answer the question why [p&q] understood in one way, by giving the following answer: I was watching it because I really wanted to see a Wim Wenders movie and you were watching it because you were dragged there by your friend.

Importantly however, if we have given this explanation, it will still seem like a coincidence that we are both watching *Kings of the Road* at the same place and time. We have given an answer to the coincidence question about our watching *Kings of the Road* that has done nothing to dispel the impression that our both watching it is a coincidence. Hence, what it is to be a coincidence has not so much to do with not being able to answer a coincidence question *at all*, but with *how* we answer it. This result gets us one step closer to understanding the placeholder concept of a *unique* explanation – the kind of explanation that is not forthcoming in cases of accidentality.

What about the way we answered the question about our movie-watching in the example above did not dispel the impression that our watching the same movie was a coincidence? In the example we have given a merely conjunctive answer to the coincidence question. That is, we have explained why I am watching *Kings of the Road and* we have explained why you are watching it, but we haven't really explained why we're *both* watching it. Or in other words, the fact that I am watching *Kings of the Road* can be explained independently from the fact that you are watching *Kings of the Road*. We haven't provided a *unified* explanation, as I will say, following Lange (2010) and Faraci (2019).

Unification can be explored further. Here is an easy way to understand why the explanation given in the example is not unified. When we ask a coincidence question in terms of why [p&q], we want to know about how p and q are *related*. We are asking a question about the relational fact that [p&q] (see Lando 2017). But the explanation which merely forms a conjunct of the explanation for p and the explanation of q does not

[131] I hereby recommend this movie to my reader.

answer this question. For all we have done is linked two separate and independent explanations for p and q – two facts which therefore count as independent, i.e. unrelated themselves. We have explained the non-relational fact that p and we have explained the non-relational fact that q, but we haven't explained how they relate. This echoes Owen's original approach. What we have effectively done is create a composite explanation of p&q, which can be divided into two independent component explanations. We have answered a coincidence question compositely, that is.

Two formal consequences of the focus of coincidence questions on relations between facts are worth accentuating.

First, if we are providing a merely composite answer to a coincidence question, then the two explanations don't interact. Let us say that X is the explanation for p and Y is the explanation for q. Then, in a composite answer, X explains p as well as X&Y does, and it doesn't explain q at all. Y explains q as well as X&Y does, and it doesn't explain p at all (see Faraci 2019, 6 and Lange 2010; 2016).

Here is how this plays out in my example: all we've done to explain why I am watching *Kings of the Road* and you are watching *Kings of the Road* is string together the explanation for why I am watching it and the explanation for why you are watching it. But this means that I can decompose the conjunctive explanation in the following way: the fact that I wanted to see *Kings of the Road* explains why I am watching *Kings of the Road*. But it doesn't explain at all why you are watching it. Moreover, the conjunctive fact that I wanted to see it and you were dragged there by your friend doesn't explain why I am watching *Kings of the Road* any better than the fact that I wanted to see it does. Likewise, the fact you were dragged to see *Kings of the Road* explains why you are watching *Kings of the Road,* but it doesn't explain why I am watching it. And again, the conjunctive explanans doesn't explain why you are watching *Kings of the Road* any better than the fact that you were dragged there does. Call this feature *conjunctive restriction* for short.

Second, and relatedly, it is characteristic of answers to coincidence questions which *do* dispel the impression of coincidence that the 'because' operator in them does not function distributively. We are familiar with operators that exhibit this behaviour, perhaps, from deontic concepts. It does not follow from the fact that it is permissible to φ and the fact that it is permissible to ψ that it is also permissible to [φ&ψ]. This is because the fact that an agent [ψ-s and φ-s] may come with additional deontic constraints due to the interaction – that is, relation – between φ-ing and ψ-ing. Likewise, it does not follow from the fact that X explains p and the fact that X explains q that X explains [p&q]. This is because it takes more to explain [p&q] as it occurs in a coincidence question than explaining p *and* explaining q. For the latter explanation is compatible with there being

no relation whatsoever between p and q. Thus, just like in the deontic case, further relational features of how p relates to q prevent the explanation from distributing in the familiar way. Call this feature *distributive failure* for short.

The distributive peculiarity of how we answer coincidence questions might come as a surprise. For it would seem that we can at least somewhat dispel a sense of coincidence by providing a common cause explanation of why [p&q]. For example, let us say that an evil scientist and Wim Wenders enthusiast has implanted devices in both of our brains compelling us to go watch *Kings of the Road* when triggered, and that he triggered them so that we are both watching it at the same time. Now, it seems that there is a common explanation for why I am watching *Kings of the Road* and you are watching it. And it is accompanied with the impression that it wasn't a coincidence after all that we are both watching *Kings of the Road*. However, this is because we are heavily primed to read the evil scientist story in a way that provides us with the relevant relational information. That is, we instinctively understand the evil scientist to have some kind of plan common to both you and me.

But the story can be told, with some effort, in a way that blocks this narrative tendency. Let us for example say that the evil scientist is host to different personalities that exist and operate without each other's knowledge. One of these personalities is a Wim Wenders fanatic, who needs people to enjoy Wenders movies. The other just wants to fill the seats of a given cinema at a given day. The first personality has implanted a chip in my brain, the second personality has implanted a chip in your brain. Both like implanting devices in people's brains, but for different reasons. Now, it seems – to me at least – like a coincidence that we are both watching *Kings of the Road,* even though the very same cause is responsible for both my watching it and your watching it. It seems like a coincidence because I have told the story in a way that eschews all relational assumptions about [p&q]. The presence of the evil scientist explains why I am watching *Kings of the Road* and it explains why you are watching *Kings of the Road*, but it does so *independently*. Hence, it provides no relational information as to [p&q].

We can also say my story explains p in one guise(Wim Wenders fanatic) and it explains q in another guise (wanting to fill the seats), but there is no relation between these guises and so no relation between p and q is established. Note that this way of representing the feature of distributive failure shows how closely it is related to the first feature, conjunctive restriction. For if we represent the scientist of my story as X, then we can see that $X_{fanatic}$ explains why I am watching *Kings of the Road*, but it doesn't explain at all why you are watching *Kings of the Road* and $X_{cinema\text{-}seats}$ explains why you are watching *Kings of the Road*, but it doesn't explain at all why I am watching it.

When the only answer we have available to a coincidence questions exhibits these features (it is conjunctively restrictive and fails to exhibit distributive failure), I will say that it is a *mere coincidence* that [p&q].[132] Mere coincidences are the result of there being no *unified* explanation of [p&q]. A unified explanation of [p&q] is an explanation of the relational fact that [p&q] (the fact that *both* facts possess some salient feature, not just that p possesses it *and* q possesses it) such that the explanation is not conjunctively restricted and it exhibits distributive failure.

As the foregoing discussion should have already shown, mere coincidence is a fragile thing. For it does not take much additional information in our answers to coincidence questions to dispel the sense that the co-instantiation of two facts is *entirely* coincidental. Take the case of the Wim Wenders obsessed scientist again. Assume that even though their multiple personalities don't know about each other, they are united in a shared appreciation of Wim Wenders movies. The movie theatre where you and I are watching *Kings of the Road* is often screening the Wim Wenders collection. This is why the first personality of the scientist sends me there for my enjoyment of Wim Wenders movies and it is also why the second personality wants to fill the seats of the cinema – to ensure the cinema continues to screen the Wim Wenders collection. Now I have given the explanation for why you are watching *Kings of the Road* and the explanation for why I am watching *Kings of the Road* a common root. And it seems to me that there is, corresponding to this root, now a sense in which it isn't *merely* a coincidence that we are both watching it. For there is now a sense in which *both* of us watching the same movie can be explained, that is, the relational fact we are asking about in coincidence questions can be explained to some degree.[133]

What emerges from this is that there will be a great variety of contextually determined senses in which it is not just a *mere* coincidence that [p&q]. The context relevant to these senses will be in large parts determined by what extent of relational information with regards to [p&q] will satisfy the purposes with which we were asking the coincidence question in the first place. Often when we ask about whether some co-instantiation was a coincidence, we are screening off background roots of the explanation for why [p&q]. For our everyday purposes, we are more interested in whether there are more proximal ways in which p and q might be related. If we are asking about both of our watching *Kings of the Road* with such more mundane expectations as context, then it will probably

[132] Riggs (2014) proposes a similarly strong condition for 'mere coincidence', according to which if A and B merely coincide, then there is 'nothing further to say' about their coincidence. They are 'completely independent'.

[133] If determinism is true, then all facts will have an ultimate causal root. It is unclear however, whether this means that they also necessarily have a common explanatory root.

not be enough to point to the fact that both personalities of the scientist are Wim Wenders fans. It will still be considered pretty coincidental relative to this context that we are both watching *Kings of the Road*. Consider instead another story: The friend who dragged you with them to the movie knew that I was going. They were going there in the hopes of pairing us up. This is a classic case in which it isn't a coincidence that I am watching *Kings of the Road* and you are watching *Kings of the Road*. It was by design that we both ended up watching the same movie at the same place and time.

A major determiner of whether or not it is a coincidence that [p&q] in some context will be the level of specificity with which the relational aspect in the coincidence question is described. For note that we can describe the fact we are both wearing red shoes either as the fact that we are both wearing scarlet 1989 sneakers or as the fact we are both wearing red shoes. And what explains the former proposition need not explain the latter.[134]

However, it will be unhelpful to go through all the determining factors for the many senses *of not a mere coincidence* that are possible. For I am here interested in accidentality as a kind of coincidence. And as I shall show presently, even when it is not a mere coincidence that [p&q], it might still be clearly accidental that [p&q]. So let us move on to accidentality now.

3. Asymmetrical Coincidence Questions

In order to discuss accidentality, I first need to address a complication about the nature of coincidence questions. Many coincidence questions are symmetrical. Successful answers to these questions will explain p and q – and they will explain them in a unified way. But there is another mode of asking coincidence questions in which answers to

[134] Bhogal's (2020) account of coincidence holds that coincidences are cases in which we can't properly explain the less specific proposition, or more precisely cases in which all we can do to explain why we are both wearing red shoes is to explain why we are both wearing scarlet 1989 sneakers.
The issue is of larger significance for my thesis (there was once a chapter dedicated to it). Specificity can be spelled out in terms of the proportion of possible worlds in which a proposition is true. The proposition we are both wearing red shoes is true in a larger proportion of possible worlds than the proposition that we are both wearing scarlet 1989 sneakers. Or in other words, the former proposition can be true in more ways than the latter (all types of red shoes worn by both you and me will make the proposition true). According to a plausible principle about explanation (see Weslake 2010; Garfinkel 1981; Wilson 1994), the specificity of the explanandum must be matched by the specificity of the explanans. If the explanans is more specific than the explanandum, then the explanation will count as, at the very least, bad – I prefer to say unsuccessful. For example (modified from Yablo 1992), let us say that we have trained a pigeon to peck at red targets. If we explain why the pigeon pecked in terms of the fact that the target was scarlet, this will be an unsuccessful explanation. For it will only explain this very specific pecking in terms of this very specific target, whereas we wanted to know (it seems) about the more unspecific/general fact that the pigeon pecked. Hence, explanations need to be modally robust, on this view. It has to be the case that the explanans holds in most (or all) of the worlds in which the explanandum holds. This is obviously a way to reintroduce modalist principles through the backdoor. My position is that when we explain actions in terms of reasons, we want to know about highly specific facts – facts about why a particular individual did what they did. Unfortunately, I had to remove the relevant chapter from the thesis in order to adhere to the mandated word count.

them do not require us to explain *both* p and q. Take for example coincidence questions about causal matters. Let us say that I ask: 'Why did the power go out when the lighting struck? Was that coincidence?' We might at first interpret this coincidence question as asking about the fact that lighting struck and the fact that the power went out. But our interests in asking questions like these are often more specific. In this case, my interest might very well lie in understanding the nature of the power outage better. When I ask why the power went out when the lighting struck, that is, I am asking a question *about the power outage*.[135]

If my question is about the power outage, then what is it asking? Plausibly, it is asking about the power outage why it 'matched', as it were, in terms of one of its properties with the lightning strike. In this case, it is plausible to assume that the matching we have in mind can be understood in terms of close temporal proximity. Why, I am wondering in my question, is the power outage such that it occurs in the same (matching) temporal segment as the lightning strike. Let us call the fact that the power went out E1 and the fact that the lightning struck E2. The question is then why E1 has the property of temporally coinciding with E2. It therefore has the familiar form of coincidence questions. But it doesn't ask about both E1 and E2. It asks why an event occurred such that it temporally coincided with E2, or: Why [E1&E1(E2)]? – where the formulation 'E1(E2)' designates that we are asking about the property of (temporally) matching with E2, not about E2 itself. I call this an *asymmetrical* coincidence question. This question behaves structurally like a coincidence question (it also asks about a relational fact), but it is not equivalent to asking why [E1&E2]. We can see this if we consider plausible answers for the case.

When isn't it a coincidence that the power outage occurs in the same timeframe as the lighting strike? Well, when the lighting strike *causes* the power outage, is one (but not the only) obvious answer. But if we were asking the symmetrical coincidence question why [E1&E2], this answer would make no sense. For it presupposes that the lighting strike occurs rather than explaining it. The answer *does* make sense when understood as answering the asymmetrical question about the power outage, however. For the answer to the coincidence question so understood may safely presuppose rather than explain the occurrence of the lighting strike. What it requires information about is the relation between the occurrence of the power outage and its temporal matching with the lighting strike.

[135] Some hold that the relata of explanatory relations cannot be particular events, but existential generalisations that range over events (most famously Davidson 1967 holds this view). On this view, asking why the power went out is not asking about the particular power outage, it is asking about why there was some event such that it was a power outage.

The form of the coincidence question under consideration (symmetrical or asymmetrical) does not change how answers to them affect our sense of whether it is a coincidence that [p&q]. It isn't a coincidence that the power outage matches with the lighting strike if the lighting strike causes the power outage, because the lighting strike causing the power outage offers a unified answer to the coincidence question. It offers paradigmatically relational information about how the fact that the power went out is related to the fact that the power going out is in the same temporal bracket as the lighting striking.

There are of course also plenty of unsuccessful answers to the coincidence question about the power outage. Imagine that the power went out because I pushed the emergency off button, and I pushed the button just slightly after the lightning struck. Now, we can explain why E1&E1(E2) by explaining why the power went off and by explaining why the power went off in the same temporal bracket as the lightning strike. But this is not a unified answer to the coincidence question. It explains E1 and it explains E1(E2), but it doesn't provide the essential relational information. Correspondingly, it has the features of conjunctive restriction (because my switching off the power explains E1 but it doesn't explain E1(E2) at all) and lack of failure of distribution (because explaining E1 and explaining E1(E2) does not entail explaining [E1&E1(E2)]).

There are again of course ways to tell this story which will dispel the sense that the power outage being in the same temporal bracket as the lighting strike is a mere coincidence. Assume for example that I was worried a very probable lightning strike would cause electrical failure and so I switched the power off. Now, it isn't a *mere* coincidence that the power went off around the same time the lightning struck.

With this out of the way, let us look at coincidence questions about intentional items like beliefs, actions, and most importantly of course responses to reasons.

4. Coincidence Questions and Accidentality

I reviewed in chapter 3 the susceptibility of orthonomy notions to coincidence intuitions. S knows that p only if S's belief that p isn't accidentally true. S's action is morally worthy only if S didn't do the right thing accidentally. And of course, S's φ-ing is a response to S's reason to φ only if S's φ-ing doesn't accidentally conform to S's reason to φ.

We can now approach these requirements from a new angle. For we can see that for each example notion, these non-accidentality requirements correspond to asking a coincidence question about the notion. For knowledge, we are asking whether the belief that p is accidentally true. This is a question about S's belief (hence, we are asking an

asymmetrical coincidence question). We are asking about that belief why it is both held by S and true. When we ask the same question for moral worth, we are asking why the agent performs the action and it conforms to the relevant deontic standard. And finally, when we ask the question about reasons-responsiveness, we are asking why the agent performs that action and it corresponds to the reason the agent has for performing it.

It is important to recall at this stage that the main marker I identified for non-coincidentality above is that we are providing an answer not only to the questions why p and why q, but an answer to the question why [p&q], which inquires about the relational fact [p&q]. This aspect, recall, echoes Owen's starting characterisation that coincidences are composite events dividable into separate components. The expanded explanationist way of understanding this characterisation is that coincidences are composite facts whose explanations can be divided into separate component explanations.

Hence, when we ask why [p&q], where p designates the fact the S φ-ed and q designates the fact that S's φ-ing corresponded to S's reason for φ-ing, we are asking about a relational fact too. Let's call the property of conforming to a reason 'rationality' for presentational convenience.[136] We can then express the coincidence question as asking not only why the action was performed and why it was rational, but why the *rational action* was performed. There is a difference, in other words, between explaining that S performed the action which happened to be the rational thing to do and explaining the rational action. The difference will be explored further throughout this part of the chapter. For now, I will express the coincidence question for reasons-responsiveness by abbreviating the fact that S φ-ed as A1 and abbreviating the fact that S's φ-ing was rational (conform to S's reason to φ) A2. Hence

Coincidence Question RR: Why [A1&A2]?

Since orthonomy notions are susceptible to coincidence questions in the indicated ways, cases of mere coincidence are familiar from the subdisciplines in which the relevant notions are situated. Textbooks and undergrad introductions on epistemology[137], for instance, are full of them. We are all familiar with cases in which an agent is struck by lightning or by a falling shingle, which causes them to believe that p. It then also

[136] When I say ‚the rational action' no theory of rationality is presupposed, and the notion does no heavy lifting beyond the presentational convenience of the term. In particular, I do not mean to commit myself to a theory o rationality according to which rationality consists in responding to reasons correctly (RRR). See Kiesewetter (2017) for a full account of RRR and a rebuttal of the arguments against such a view.
The view developed in chapter 5 of this thesis can be developed into what I think is a superior version of an RRR view – the view that rationality consists in *exercising* the capacity to respond to reasons.

[137] Cases also often feature 'a lucky guess', such as in Audi (1997), 23

happens to be the case that p is true.[138] This is a case of mere coincidence because we can only give a composite answer to the question why S believes that p and S's belief that p is true. That is, we can explain why S believes that p (by mentioning the shingle) *and* we can explain why S's belief is true (by whatever theory of truth we chose presumably). But we can't explain the *true belief*, as it were. These cases are most famous in epistemology, but they can be easily reproduced for other orthonomy notions such as RR. Just assume that I have recognized a sufficient reason to φ when suddenly a shingle hits my head and makes me φ. Again, we can explain the fact that I φ and we can explain the fact that my φ-ing conforms to a reason I recognized. But we can't explain them together. We can't explain my rational φ-ing because of course my φ-ing is not a rational φ-ing in this case. It is merely a φ-ing that happens to be also the rational thing to do. For some foreshadowing, call the devices in our stories (the shingle) that separate the two explanations for p and q *exploits*. Exploits are ways in which we can make both conjuncts of a conjunction true without establishing any connection between them. They are thereby exploiting overly composite treatments of orthonomy notions, such as for example:

Silly Knowledge Composite: S knows that p iff S believes that p *and* p is true (and justified).

Exploit cases are familiar from their subdisciplines, because they are in part what motivated early causalist approaches to orthonomy.[139] Recall that one easy way to dispel the impression that it is a coincidence that [E1&E1(E2)] is to point out that E2 caused E1. The basic causalist intuition about orthonomy is that the same type of answer will work for responding to reasons and knowledge and whatever other orthonomy notion we are focusing on. That is, according to the causalist, it will be enough for the agent to be in orthonomous relationship with the world if the relevant aspects of the world *cause* the agents mental states or actions.

But we already know that causalism faces a deep problem. For we know that the causal answers that work for coincidence questions about events (and other non-intentional items) do not work for intentional items like beliefs and actions. We know this because it

[138] Gettier-cases, famously introduced by Gettier (1963) are, with some caveats, also of this sort. If I look at a clock that is stuck on 20.00 and on the basis of this reading form the belief that it is 20.00 while it in fact happens to be 20.00, then the stuck clock explains my belief and the fats about the time explain my belief matches with the truth. But we can't give a unified explanation of my true belief (cases like this already appear in Russell 1948)
Things are a bit more complicated with the typical Barn cases, in which the agent just happens to be looking at a real barn in a sea of fake barns. Here, it would seem that we can give a unified explanation of the agent's true belief, because the luckiness in question is, as I said in chapter 3, footnote 1 *preselectional*. In short, my view is that as long as the agent exercises their capacity to recognize real barns, they do indeed have knowledge. Knowledge is precarious according to the explanationist, after all.

[139] Such as Dretske and Enc (1984); Grice (1962); Goldman (1967); Stampe 1977.

is precisely what cases of causal deviance show. For in cases of causal deviance the right items in the world cause the agent's mental states or actions. But it is still accidental that the agent's mental states or actions conform the relevant aspects in the world. This is why we need to treat accidentality separately from coincidence. It is a kind of coincidence, but a special kind. Even when it isn't a mere coincidence that [p&q], it might still be accidental that [p&q]. As long as we can't capture what goes on in cases of deviance, we therefore cannot develop a theory of non-accidentality. To unpack this, let me briefly discuss cases of deviance, in particular as seen as instance of accidentality, in more depth.

5. Deviance and Accidentality

We have encountered deviance several times in this thesis, and each time cases of deviant explanatory chains[140] have played a significant role in my argument. In chapter 2, I argued that possession accounts of reasons-responsiveness fail because they fail to account for the fact that actions brought about deviantly are not free. In chapter 5, we saw that cases of explanatory deviance like that of Lord's Fortunate Consequent Affirmer naturally support exercise-views about what it is to have (or 'possess') and act for reasons. Let us now take a more comprehensive look at these cases.

The problem of deviant causal chains first gained notoriety within the causal theory of action (CTA).[141] The standard version of the CTA pursues the reductive aspiration of giving an analysis of what distinguishes mere behaviour from action in terms of the causal sources of behaviour. It is therefore a paradigm instance of what I called causalism above. Only behaviour caused by certain rationalising mental items counts as action according to this idea. In Davidson's famous version of the CTA, so-called belief-desire pairs (or their neural realisers) must cause behaviour for it to count as an action.

However, the Davidsonian CTA soon ran into a problem, now immortalised in the following case:

> 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 that he never chose to loosen his hold, nor did he do it intentionally. (Davidson 1973, 79)

[140] Mayr (2011), ch.5 contains a very helpful overlook over various responses to the problem of deviance.

[141] Versions of the CTA are advanced by Bishop (1989); Brand (1984); Davidson (1980); Enc (2003); Goldman (1970); Mele (1992b).

In this case, there is a belief-desire pair which rationalizes what the agent does, and this belief-desire pair causes what the agent does. But what the agent does, does not appear to be done for that reason, nor indeed is it an action at all. Something in the way the agent's doing is caused seems to have gone awry. The causal sequence leading up to her behaviour is deviant or wayward. This is the problem of causal deviance.[142]

The literature sometimes distinguishes between primary and secondary (or antecendential and consequential (Brand 1984, 18) deviance, depending on which step in the causal sequence is infested with waywardness. But the distinction is not massively important. Where exactly these possibilities of deviance will occur depends on more precise theories of practical reasoning and action that I will, for the most part, not take a stand on. If there is a solution to the problem of deviance, the solution will have to be general anyway. Here, I will presuppose cases of deviance that, like the climber-case, run between mental state and behaviour, or reason and action in cases to be discussed soon.

The problem of causal deviance poses an especially pressing problem to versions of the CTA which have reductive aspirations. For Davidson, it seems, the rationalisation and causal explanation of the behavioural token by the belief-desire pair is supposed to be all that is needed for the behavioural token to count as an action. Cases of causal deviance show that Davidson's ingredients don't add up to actions and so the reduction project is in jeopardy.

But the perspective of this thesis is that deviance is not only a problem for the reductive Davidsonian project. It is a problem for any analysis of an orthonomy notion according to which the relevant orthonomous relationship is (merely) causal (see Enc 2003, 99). Notions infested with the problem therefore include all famous orthonomy notions, such as knowledge[143], perception[144], and responding to reasons, but also more niche concepts such as rule-following[145]. This is because what unifies these notions is that they require a non-accidental connection between their items (as discussed in ch.3). And the problem of deviance, at base, is a problem of accidentality. Aguilar puts this succinctly:

> Even a perfunctory look at the examples offered in the literature to illustrate the problem of basic causal deviance reveals a key feature shared by all of them. This feature is the presence of an accidental event or sequence of events that

[142] For important treatments of the problem, see Peacocke (1979a/b), Brand (1984), Bishop (1989), Mele (2003), Schlosser (2007, 2011), Wu (2016).

[143] See Pollock and Cruz (1999).

[144] See Peacocke (1979a), 123.

[145] See Schlosser (2011).

> causally links an internal motivating state with its intended outcome. That is, in each case some fortuitous occurrence takes place linking the intended outcome with the motivating internal state that starts the causal chain and whose content is precisely the one matched by the outcome. (Aguilar 2012, 3)

In accordance with the explanationist approach I have taken to coincidence, we can put this insight in explanatory terms. The root of the problem lies in the fact that the mere causal connection between items fails to amount to the right type of explanation of – in Davidson's case – the behaviour. We can see this sort of structure at work clearly in causal explanation approaches to 'acting for a reason'. In chapter 5 I discussed, and agreed with, the view that acting for reasons encodes reasons-explanations of actions that are a kind of causal explanation. According to these widespread views, to say that S φ-ed for some reason p is to say that there is a true causal explanation of S's φ-ing available in terms of p. So this is a theory that characterizes acting for a reason in terms of the reason figuring in the explanation of the action. But now consider Davidson's climber again. In the climber case, the climber certainly does not act for the reason provided by (or in) the desire (and corresponding instrumental belief) to let go. And yet it seems that this belief-desire pair is involved in the explanation of the behaviour i.e. the belief-desire pair certainly causally explains the behavioural token in this case. So characterising acting for reasons in terms of true causal explanations that mention the right reasons does not seem to be enough, because the reason may explain the action in the wrong way – in a way that does not disperse the impression of accidentality. This is also why I have been emphasising that deviance cases are problematic because they feature the wrong type of *explanation*, not just wayward causation. What is going on metaphysically in the climber case is that there is a deviant causal process leading from the belief-desire pair to the behaviour. But it seems to me that the more important aspect, for now, is that this corresponds to a type of causal explanation that for some reason is not eligible for grounding the relevant reasons-explanation (the relation between causation and causal explanation is discussed in the next chapter so for now, I will put things mainly in explanatory terms).

It is worth dwelling on this point a bit more. For you might be sceptical about my claim that in the climber case the belief-desire pair *does* causally explain. The very fact that a true reasons-explanation is not forthcoming the climber case, you might say, shows that the belief-desire pair is not what causally explains the behaviour. But I suspect that this is the same point I just made, only put in less helpful terms. For you might defend your claim by pointing out that in the climber case, the belief-desire pair did not causally explain qua its content, or qua its rationalising role (see Schlosser 2007) or more generally, that in deviant cases reasons don't explain qua reasons. But with this I agree.

Making this point doesn't deny however that there is a causal explanation in terms of the belief-desire pair under some description where that pair is the efficient cause. I agree that this explanation will not be very informative for most contexts, as it just provides a mostly irrelevant snapshot of the causal history of the relevant behaviour. But nevertheless, it will provide some explanation of the behaviour in terms of its causal history, and so an explanation it will still be. It is clear that what is needed for solving the problem of deviance is a narrower sense of explanation on the basis of which we are able to distinguish unfitting explanations like the one that remains true in the climber case from fitting explanations like those that correspond to non-deviant production of items in the 'normal' cases. But saying that this special class of explanations is demarcated by those cases in which reasons explain qua reasons doesn't really move us any closer towards a helpful general demarcation criterion. It just reiterates the general lesson of deviance cases: those causal explanations on which we aim to ground some of our most important concepts in orthonomy accounts are somehow special. What we need to find is an account of their specialness. In fact, if we find the (generalisable) sense in which it is untrue that the belief-desire pair explains the behaviour in the climber case, we will have ipso facto found the missing explanatory property we started looking for, the property that the explanationist thinks illuminates non-accidentality. And my view of coincidence developed in the previous section opens up a route towards that property.

6. An Explanationist Account of Non-Accidentality Part 1

My explanationist approach to coincidence discussed above offers a new way of understanding the problem of deviance, a way that allows us to construct a structural solution to it and therefore to find the relevant explanatory property – the property missing in cases of accidentality. For it allows us to see that the problem of deviance is just a special case of the problem structure we already identified. It is a case of answering the relevant coincidence question compositely, that is.

To see this, consider again Davidson's climber example discussed above. Take note especially of how Davidson's underlying version of the CTA conceptualises *action*. According to Davidson, some behaviour counts as an action iff what causes that behaviour is also what makes it rational (which for Davidson, is a belief-desire pair). Thus, Davidson holds:

Davidson Composite: b-ing is an action iff m causes b-ing and m rationalises b-ing.

As a consequence, the Davidsonian account can only give an unsatisfactory answer to the coincidence question about actions. The relevant coincidence question is: Why did

S b and S's b-ing conformed to what S intended (I use intention as equivalent to belief-desire pair)? Was it an accident that S's behaviour conformed to S's intentions?

If we are committed to an analysis like Davisonian Composite, our answer can only be the following: m caused S's b-ing and m rationalised S's b-ing. But this is evidently a composite answer to the coincidence question. For m explains why S b-ed, and m explains why S's b-ing was rational. But m doesn't explain S's rational b-ing. This is because m plays two independent explanatory roles: that of a causer of S's b-ing and that of a rationaliser of S's b-ing. But it plays those roles independently. And so it does not deliver the relational information the question why [p&q] requires.

Cases of causal deviance exemplify the remaining possibility of this composite way of answering the coincidence question. For the stories they tell employ exploits that make sure to separate m's dual explanatory role. In the climber case, the climber's intention makes the climber so nervous that they let go. Thus, the nervousness separates the causal role of the intention from its rationalising role. The causal aspect of the intention still causally explains why the climber let go. And the rationalising aspect of the intention still explains why the climber's behaviour conforms to what they intended to do. But the coincidence question is asking for how these two aspects are related, and for all we know from Davidsonian Composite, they are not related at all.

I submit that all cases of accidentality, certainly all cases of deviance can (and should) be analysed along these lines. Pertinent to my interests are of course the cases of deviance discussed in chapter 5 and briefly in chapter 3. Recall that a decisive case in chapter 5 was that of the Fortunate Consequent Affirmer. Here it is again, as Lord & Sylvan present it:

> Fortunate Consequent-Affirmer. Sam wonders whether Terry took the bus to work. He knows that Terry's car is in the driveway. This is, in fact, a sufficient abductive reason to think that Terry took the bus. Sam also believes that if Terry took the bus, then Terry's car is in the driveway. But he comes to believe that Terry took the bus by inferring that he took the bus from his own belief that Terry's car is in the driveway and his belief that if Terry took the bus, then Terry's car is in the driveway by following an invalid deductive rule: from <if A then B>, and <B>, infer <A>. Sam hereby manifests a general consequent-affirming incompetence. (Lord and Sylvan 2019, 148)

In chapter 5, this case was important because it establishes that accidental success is worse than non-accidental failure. That is, I used it to back up the view that the line between responsive and unresponsive agents is best drawn using the notion of an exercise of the capacity to respond to reasons.

Looking at the case with the apparatus of this chapter, we can see that Sam's belief accidentally conforms to a reason because we can only give a composite, non-unified answer to the relevant coincidence question.

We are asking: Why does Sam believe that Terry took the bus to work and this belief conforms to the reason for believing that Terry took the bus to work. But in Sam's case, the only answer we can give is the following: The fact that Terry's car is in the driveway explains why Sam believes that Terry took the bus to work (through the weird reasoning rule) and the fact that Terry's car is in the driveway explains why Sam's belief conforms to his reason. But because Sam arrives at his belief through the strange reasoning rule, these two explanations can be divided into two separate components. The fact that Terry's car is in the driveway plays two independent explanatory roles. It explains why Sam holds the belief and it explains why it is rational to hold the belief. But it doesn't provide the required relational information. Hence, it is still accidental that Sam's belief matches his reason.

Just like for Davidson, the case exists because of an underlying composite treatment of the responding relation. That is, the case is the logical consequence of treating responding to reasons as follows:

RR Composite: S's φ-ing is a response to the reason R iff R causes S's φ-ing and R rationalises S's φ-ing.

In order for some φ-ing to count as the agent's responding to reasons for and against φ-ing, we need a unified explanation of 'the rational action'. That is, we need an answer to the coincidence question for reasons-responsiveness: Why [A1&A2]? In cases of deviance, all we can do is explain A1 and A2 separately. We can explain why the agent φ-s (A1) and we can explain why the agent's φ-ing matches their reason (A2). But we can't explain why it is a reasons-matching φ-ing (the formulation is supposed to highlight the relational reading of '[A1&A2]'). More precisely, the way in which these explanations work is that the same proposition plays two independent explanatory roles in them. As a cause, the proposition explains the occurrence of the agent's action. As a rationaliser – a consideration that makes it rational to φ – the proposition explains why the φ-ing is rational. But, again, this is a composite answer to the coincidence question, which does not dispel the impression that it is accidental that the agent's φ-ing matches their reasons (that they do what their reasons recommend). What is needed to dispel the impression of accidentality is relational information as to how the fact that the proposition caused the action is related to the fact that it is what rationalises it.

This last formulation rejuvenates an old way to look at cases of deviance. According to this old way, the problem of deviance is about synthesising two different relations: A

purely causal relation and the rationalisation relation. When reasons explain actions (in the right way) – when agents act for reasons, that is –, these two relations have to be, in some sense *the same*. Reasons '*ratio-cause*' actions, as it were. My treatment gives a more precise expression to this idea. It holds that cases of deviance two independent explanations of two independent explananda exist when there should be a single non-composite explanation of one non-dividable relational fact. In structural terms:

Accidentality Structure: Why [A1&A2]? (A1 because R_{cause}) & (A2 because $R_{rationalise}$)

This approach tells us exactly in what sense cases of accidentality (of which deviance is a type) can be understood as involving an explanatory defect. But I wanted an account of *non*-accidentality – the kind of explanatory 'virtue' that has to obtain on order for relationships to be non-accidental. In order to get to such an account, we need to follow the breadcrumbs left by my treatment of deviance just a bit more. This last step is taken in the next section.

7. An Explanationist Account of Non-Accidentality Part 2

Now we know what goes wrong explanatorily in cases of accidentality. But we need to also know what goes *right* in cases of *non*-accidentality. What type of answer to the coincidence question for RR will dispel the impression of accidentality? The foregoing discussion has already provided a rough blueprint for the answer. The central failure that occurs in deviance cases is that we offer two separate explanations for two separate explananda. Consequently, we fail to provide the relational information the question is inquiring about.

Therefore, a successful (i.e. accidentality-dispelling) answer to the coincidence question will somehow have to speak about the relation of the fact that S φ-ed and the fact that S's φ-ing is rational. In fact, we intuitively know what the relation of these facts will look like for non-accidental relationships between reason and action. When agents are responding to reasons, then these reasons cause their actions *because* they rationalise them.[146] This is after all the idea behind orthonomy, that agents are moved by what they recognise to be good reasons for being so moved. Hence, we should expect cases of non-accidentality to follow a structure like:

[146] Lord (2018), 174 pursues the same intuition, spelling out a dispositional 'in virtue of' relation.

Expected Non-Accidentality Structure:

Why [A1&A2]? Explanation1: (A1 because$_{cause}$ R) & Explanation2: (A2 because$_{rationalise}$ R) & (E1 because E2)

Unfortunately, this solution still does not deliver the right structure for non-accidentality for the same structural problem I identified above. It still constitutes a composite answer to the coincidence question.

To see this, notice that Expected Non-Accidentality Structure makes use of an unsubscripted and somewhat mysterious explanatory relation in the third conjunct. This is a problem. For recall that it is true *in a sense* in cases of deviance that the reason *explains* the agent's φ-ing. It is just not true in the right sense – the sense that would make the impression of accidentality disappear. The issue with linking the causal role and the rationalising role of the reason explanatorily is that *this* explanatory connection is subject to the very same problem. There are many ways in which any sentence 'p because q' might be true, only some of which will give us the sense we require in order to dispel the impression of accidentality.

The literature on deviance is already populated by examples that manifest precisely this kind of structure in fact (see Mantel 2018, ch.2; Bishop 1989, 151; Stout 2010, 163). Imagine for example that the nervousness in the climber case is replaced by a highly complex machine. The machine will register the precise contents of the climber's motivational states and is fitted with an algorithm that will make the climber produce precisely those behaviours that fit those states. Thus, the rationalising role of the climber's desires will, in a sense, explain its causal role. But because the explanation includes the weird device, it will not be the sense required for excluding accidentality. In the case of the climber with the device, when the device is active, the climber's behaviour still does not count as an action after all. It doesn't count as an action because it still seems accidental that that the climber does what they desired to do. The rationalising role of the desire explains its causal role in the sense that it fits with the causal role (the causal outcome match the rationalised behaviour) and in the sense that it brings about the causal role (it triggers it, as it were), but it plays these roles independently, and so again a composite structure emerges. Again, the answer to the coincidence question is ultimately composite in nature.

This was a case for action which illustrates the problem well. But of course we can easily generate the same structure for cases of responsiveness to reasons. Imagine for instance Sam from the Fortunate Consequent Affirmer case, whose consequent-affirming disposition has been replaced by a device implanted in his brain. The device is programmed to pick up on sufficient reasons for beliefs and induce the appropriate

beliefs in the agent. And when Sam discovers that Terry's car is in the driveway, this is exactly what the device does. It induces in Sam the belief that Terry took the bus to work.

It seems that Sam's belief still accidentally fits his reason. Sam believes what his reasons recommend accidentally. So Sam does not respond to this reason. But his reason causes and makes rational his belief, and it causes the belief because it makes it rational (due to the device).

A slight complication at this stage necessitates a detour.

If you are still haunted by modalist intuitions, you might have a different opinion about these examples. You might think that the presence of a translational device makes the connection between reason and belief (in the Fortunate Consequent Affirmer case) or action modally robust. Hence, you might then latch onto a sense in which in these examples Sam's belief doesn't accidentally match with the reason, the climber's action doesn't just accidentally match their intention.

There is indeed an aspect to these cases that confuses intuitions. But it isn't the fact that the connection between the relevant items is modally robust. Think about the following case. Due to being dropped by their buddy some years ago, a climber has a dead arm with no motor control. However, a device which sits in their spine detects intentions and translates them into the corresponding arm movements. Here, it is at the very least unclear if the climber's movements count as accidentally fitting their intentions. Our intuitions here are guided by the extent to which the deviant aspect is *functionally integrated*. Prosthetic limbs or sophisticated electronic replacements of damaged nerve connections are, in a sense, deviant connections (see Bishop 1989, 159 and Mayr 2011, 118 for similar cases).[147] But we are often unsure about whether these connections pose any philosophical trouble. This is because they feature translational devices that are fully functionally integrated within the system in question.

We can disentangle intuitions about full functional integration and modal robustness. For we can run the new deviance + device cases above by explicitly stipulating away the assumption that relevant exploits are functionally integrated. I submit that with this stipulation in place, we can't deny that even though the connection between reason and

[147] The debate is also sometimes framed in terms of whether *other agents* can be involved in non-deviant chains. Dretske (1992) is puzzled by the fact that when we make other people do things by presenting them with reasons, they act as intermediary agents for our intentions. He begins his curious paper on the phenomenon by saying:

> I offer Jimmy a dollar to wiggle his ears. He wiggles them because he wants the dollar and, as a result of my offer, thinks he will earn it by wiggling his ears. So I cause him to believe something that explains, or helps to explain, why he wiggles his ears. If I push a button, and a bell, wired to the button, rings because the button is depressed, I cause the bell to ring. I make it ring. Indeed, I ring it. So why don't I, by offering him a dollar, make Jimmy wiggle his ears? Why, indeed, don't I wiggle them? If I ring a bell by pushing a button, why don't I wiggle Jimmy's ears by offering him a dollar? (1)

belief/action is modally robust, it is still in the relevant sense an accident that the belief/action matches with the reason.

This result is telling for the overall narrative of this thesis. For it turns out that the deviance + device cases are already familiar to us. They are structurally identical to cases of (dispositional) mimics (which belong to what I called 'modal exploits'). The device, when not fully functionally integrated, mimics the agent's reasons-responsiveness just like a disintegration device linked to a rock and set to trigger should the rock hit a hard object mimics fragility. What this shows is that there is a link between cases of finking, masking, and mimicking and the other cases of exploits of composite structures reviewed in this chapter. Finking, masking, and mimicking are cases of modal accidentality, that is. I will further develop this important result in section 8.

End of detour. The lesson from the climber + device case (and cases like it) should then be this: the desire causes the behaviour, the desire rationalises the behaviour and the desire causes the behaviour because it rationalises it (via the device). But the behaviour *still* isn't an action. This is because due to the device, although it is true there is a sense of 'because' in which the desire causes because it rationalises, this is evidently not the right sense of because. The same goes for the RR case, mutatis mutandis.

Think about why we have this impression. It seems like giving the type of answer to the relevant coincidence question that involves a kind of translational device does not provide the relational information we are seeking. Despite the presence of the device, we are still dealing with two separate ways in which a proposition is related to an action/belief. That is, even though rationalising role and causal role are connected by the device, it still seems like the reason (or desire) independently rationalises and causes the action. So we are still explaining A1 in virtue of the causal role and A2 in virtue of the rationalising role of the desire/reason. The root of the problem lies in the explanatory relationship between the causal and the rationalising role of the reason. For although the fact that the reason plays a causal role is explained by the fact that it plays a rationalising role, it is not explained by that fact in the right way. This is because the explanation is itself decomposable into independent components in the same way in which answers to the coincidence questions are in the original cases of deviance. The best way to see this is to realise that we can give two descriptions of the explanation of the fact that R causes Sam's belief in the deviance + device example.

On one description, we explain why the reason, *considered as the trigger of the device*, plays a causal role. On the other, we explain why the reason, *considered as a normative consideration*, plays a causal role. We can split the explanation for why the reason plays a causal role in this way, because we can split the explanans, i.e. the fact that the reason

is playing a rationalising role into two more fine-grained descriptions. Further, the possibility of describing the rationalising role in two more fine-grained ways goes along with the possibility of doing the same with the causal role (i.e. the fact that the reason causes the action). For we can either describe the reason as a causal factor in virtue of the device, or as a causal factor in virtue of the normative force the reason has.

Thus, structurally, the last conjunct of Expected Non-Accidentality Structure looks like this (notice the italicised part)

Expected Non-Accidentality Structure*:

Why [A1&A2]?	E1: (A1 because$_{cause}$ R) & E2: (A2 because$_{rationalise}$ R) & *((R_{cause1} because $R_{rationalise1}$) & (R_{cause2} because $R_{rationalise2}$))*

Thus, in the end, Expected Non-Accidentality Structure still features a composite i.e. non-unified answer to the relevant coincidence question.

The root of the problem lies in the fact that the explanation we require to disperse the impression of accidentality is highly sensitive to minute changes in the description of the proposition in explanans position, i.e. the reasons providing proposition.[148] The explanation is so sensitive to differences in description, in fact, that two (or more) unrelated more specific descriptions of the explanans proposition will lead to the decomposability of the explanation into two (or more) independent sub-explanations in terms of the more specific description of the proposition.

I am assuming here that each description *I* of the proposition corresponds to a specific explanatory role. And for each explanatory connection we posit between the fact that the proposition plays role 1 (corresponding to description *i*1) and the fact that the proposition plays role 2 (corresponding to description *i*2), we get even more specific descriptions of those roles 1 and 2. For each of these roles, if they are unconnected, we get our own sub-explanation corresponding to the specific role and description

[148] If you are sceptical of the notion that the same proposition can be given different descriptions, you can still accept what I am proposing here. In fact, I agree that a better way to talk about the current issue is in terms of what Daniel Nolan calls 'hyperintensional metaphysics' (Nolan 2014). The idea here is that propositions can be individuated in a highly fine-grained way, so there isn't one reasons proposition under different descriptions but different hyperintensionally individuated reasons-propositions, which need to be explanatorily connected. I have found, though, that presenting the issue in this way has proven more distracting to the point I am trying to make than the version which focusses on descriptions, probably because philosophers still take hyperintensional metaphysics to be a bogeyman.

assigned. And as soon as a sub-explanation is itself decomposable, we get the possibility of accidentality.

At this point it becomes obvious that the strategy of adding explanatory connections to our answer to the coincidence question faces a regress. For each added connection, we will have to block the decomposability of that explanation by way of adding another connection, which we then have to make 'undecomposable' by adding another connection and so on…until what?

The regress is not infinite. The possibility of splitting up an explanation in the way represented in Non-Accidentality Structure* depends on the possibility of assigning more and more fine-grained explanatory roles to R. And such graining isn't endless. There is a point at which our resources of description will not permit a further splitting some role of R into two further roles. We know that the limit here is not extensional equivalence of R_1 and R_2, because we are evidently dealing with a hyperintensional phenomenon. That is, it is possible that there is a true explanation in virtue of R_1 that is not true when we replace R_1 with R_2 even though R_1 and R_2 are co-intensional.

I can't go into the semantic resources of hyperintensionality for reasons of space. But it will not matter much here what the limit of grain is exactly, as long as we see that it must exist. We can then use this limit, wherever it lies, to understand non-accidentality. For what we have now found is that deviance cases not only show that the answers to coincidence questions have to be unified (as they have to be for coincidence-dispelling answers), they have to be *maximally unified*. An answer to a coincidence question is maximally unified iff *none* of its parts is decomposable into separate independent explanations. Or, expressed in terms of explanatory roles:

Maximal Unification RR: The explanantia R in an answer to an coincidence question are maximally unified iff there is no description of R under which R can play a sperate independent explanatory role (i.e. all roles R can play are themselves connected by a unified explanation).

The purpose of this chapter is to find the explanatory property that sets accidentality-dispelling explanations of [A1&A2] apart from explanations of [A1&A2] compatible with it being accidental that [A1&A2]. This type of explanation is what is not forthcoming in cases of deviance. Maximal Unification RR spells out this property (for RR notions). In order for it to be non-accidental that the agent φ-ed and their φ-ing conformed to the sufficient reason R for φ-ing, R must explain [A1&A2] in a maximally unified way. That is, R must explain [A1&A2] such that there is no way to decompose '[A1&A2] because R'

into independent explanations. In other words, if we want to explain the rational action in a accidentality-dispelling way, then we need to explain it such that there is no possibility that we are ultimately only explaining why the agent φ-ed and it was rational for them to φ. In order not to *ultimately* explain the non-relational circumstances, it is not enough to just establish some connection between the explanation of A1 and the explanation for A2. For if this connection is itself decomposable, the phenomenon of accidentality will reoccur. What we need to provide is therefore an explanation that delivers relational information for all possible ways in which we can understand R.

Thus, Maximal Unification finally allows us to formulate an explanationist account of non-accidentality:

Non-Accidentality Structure RR**:

Why [A1&A2]?	[A1&A2] because$_{\text{maximally unified}}$ R
Non-Accidentality (RR):	It is non-accidental that [A1&A2] iff R explains$_{\text{maximally unified}}$ why [A1&A2][149]

This account tells us what has to be true about an action or attitude if it is to count as non-accidentally connected to a reason. What has to be true is that there is an explanation available in terms of that reason which gives a maximally unified answer to the question why the agent's φ-ing matched with her reason – a maximally unified explanation of the rational action, that is. Thus, the account understands the crucial non-accidental relationship between reason and action in terms of the special way in which the reason explains the action when it explains why it matched with the reason.

To be clear: What the account thereby offers is an understanding of non-accidentality in terms of a reasons-explanation of action. Reasons-explanations of action explain the agent's rational action in terms of a reason, after all. This insight requires us to look at the form of reasons-explanations of action in a new (or deeper) way.

First of all, my treatment allows us to see that reasons-explanations have a special relational explanandum: the relational fact that the action occurred and matched with the agent's reason. This insight, properly interpreted, requires us to think in a different way about actions. There is a tendency still in the theory of action to think about actions as non-relational items, paradigmatically events.[150] But as I shall explore more in the next chapter, explaining the relational fact that S's φ-ing matched with their reason R will

[149] The principles are already formulated with RR in mind, but it should be obvious how they generalise for all orthonomy notions.

[150] The view is so widespread that Bach (1980) calls it a "common prejudice".

mean explaining why R caused S to φ. Reasons-explanations of actions then explain causings rather than the effects of these causings (see ch.7 section 4).

Second, we have to recognize that the explanans of reasons-explanations is more complicated than is often assumed. It isn't identical, for example, with the reasons-proposition, i.e. the fact which provides the relevant reason. For as deviance cases show, this proposition is not enough to explain the relevant relational aspects in our question. The next chapter holds that what is missing is the presence of a dispositonal property in the explanation – the capacity to respond to reasons.

But even without these two points fully spelled out, the view helps understand what it is for an action/attitude to be a response to a reason:

Explanationist Responding: S's φ-ing is a response to S's reason R to φ iff R explains$_{\text{maximally unified}}$ why [A1&A2]

Note that no counterfactualist or possibilist vocabulary is required in this structural understanding of non-accidentality in explanationist terms. All we need is to understand the explanatory demands of the coincidence question and the types of answer that meet them. We do not need to have any information, that is, about what would have happened in slightly different possibilities.

In fact, the idea that a relationship is non-accidental only if we can establish the right sort of explanatory connection affords us a way to see that the modalist treatment of non-accidentality cannot possibly succeed. The remainder of this chapter is dedicated to presenting this argument.

8. Why Modalism Fails

Recall that the account developed in this chapter is meant to rival the modalist treatments non-accidental relationships that are usually given. Modalism holds roughly that non-accidental relationships are (or can be understood in terms of) relations of modal tracking of facts across modal subspaces.

Here is what I take to be the best general argument against any sort of modalist analysis of non-accidental relationships in a nutshell: throughout this paper, what we've learned is that non-accidental relationships cannot be composites, decomposable into separate independent parts. Modalism however must treat non-accidental relationships as exactly that. It treats them as modal composites, decomposable into separate independent truths across a range of worlds. Composite treatments of non-accidental relationships, we have learned, always give rise to exploits of some sort – stories we can tell that bring

to mind the possibility that there might be no relation between the component parts of a composite. And because modalism is a composite account of non-accidental relationships, it comes with exploits too, which I have called 'modal exploits' in earlier chapters. Modal exploits are stories we can tell that bring to mind the possibility that there might be no relation between the modal components of the modalist composite. These are the cases we already know – they are finks, masks, and mimics, to which, most importantly, Frankfurt-like cases in all subdisciplines belong.

This was the argument (and its systematizing potential) in a nutshell. Now let us look at it in more detail. Let us start with zooming out a little bit to the spirit of the explanationist account of non-accidentality developed in this chapter. The core of the account is that facts about non-accidental connections cannot be decomposed. I happen to think that this insight is best captured in explanatory terms as the view that to explain a fact about a non-accidental connection is to explain a relational fact in a unified way. But the principle that underlies this view, which it shares with traditionalist accounts, is a principle about decomposability. Recall that traditionalist views hold that coincidences are co-instantiated, independent events with no common cause. Non-coincidentally related events consequently are events that do share a connection (for example in the form of a common cause). Consequently, we cannot decompose facts about these events into two (or more) separate non-relational facts. Facts about non-coincidentally related events *must* be understood as not decomposable then. As long as we can ultimately reduce them to two or more non-relational separate facts, there will be a corresponding sense of coincidence there. This is especially true for non-accidentality. Our intuitions for this special sort of coincidence are centrally guided by whether the relevant facts decompose or not. The fact that S believes that p and S's belief matches with p (i.e. is true) is a fact about S's knowledge of p only if we cannot decompose it into the separate non-relational facts that S believes that p and that S's belief matches with p. The fact that S φ-s and S's φ-ing matches with S's reason R to φ is a fact about S's responding to R only if we cannot decompose it in to the independent non-relational fact that S φ-s and S's φ-ing matches with R. We know, moreover, that even relational information about R will often not be enough to instil a sense of non-accidentality. This is because the fact that S φ-s and S's φ-ing matches with R counts as decomposable when there are separate, independent descriptions of R. It will then decompose into separate, independent more fine-grained facts that involve the respective description of R. What we express when we say the rational action was performed is not a decomposable fact like that. I.e. we do not merely express the fact that S φ-ed and S's φ-ing matched with R1 (for example R1 under a causal description) *and* S φ-ed and S's φ-ing matched with R (for example under a rationalising description). That is, we are not just expressing the

thought that S's φ-ing was caused and rationalised independently by R. We express the unitary fact that R did both.

We can then record that there is a fundamental decomposability principle at work in or intuitions about accidentality and coincidence:

Decomposability Principle: If we can decompose the fact that [p&q] into two separate, independent facts p and q, there will remain a sense in which it is accidental that [p&q].

My point is that modalism, by its very nature, clashes with this fundamental principle and therefore *cannot* succeed. That is, the way in which modalism analyses facts about non-accidentality makes them ultimately decomposable, and thereby accidental by the Decomposability Principle.

To see this, we need to also assess the spirit of the modalist account briefly. Recall my representative for modalism:

Modalist The fact that p is non-accidentally connected to the fact that q iff a sufficient proportion of the relevant p-worlds are q-worlds.

For reasons-responsiveness, this translates into the now familiar view that to respond to a reason means that the agent acts according to similar reasons in a sufficient proportion of relevantly similar worlds, i.e. that the agent φ-s and her φ-ing matches her reason to φ in a sufficient proportion of the relevant worlds.

Modalism is a kind of pattern view. We are to imagine the actual non-accidental connection between p and q as emerging from a specific arrangement of facts in the mosaic of alternative possibilities. If the two facts pattern erratically across the relevant subspace, no clear arrangement will emerge, and so no non-accidental relationship can be assumed between them in actuality. They need to be 'companions', in other words: coinstantiated in most or all of the relevant worlds.[151]

This way of presenting the core modalist thought perhaps already highlights the problem. Patterns, as we know, may accidentally amount to a given supervening picture. Imagine a board fitted with pixel-like slots, each slot holding two lights. Now imagine that across some subspace of that board (say a square area in the centre) these lights light up together. What, if anything, does this fact tell us about the relation between these lights? It tells us that they are companions across a subspace of the board, and that is it. For it is entirely possible that each light is controlled by its own circuit, and it

[151] I take the companions metaphor from Felipe and Tognazzini (2010).

might have been the case that each light switch for the lights in the defined area was flipped on some whim. If this is the case, then there is no significant relation between the lights. The fact that they pattern across a subspace is a coincidence – it is merely the composite of the fat that light 1 lit up, light 2 lit up and so on. The story about the light switches being flipped on a whim is an exploit for the example. It makes the relevant truths (about light 1, 2..) obtain without establishing any relation between them.

Modalism has essentially the same problem. It understands non-accidental relationships between p and q in terms of p and q being companions across the relevant subspace. Thus, it treats the fact that it is non-accidental that [p&q] as a composite of a range of independent separate modal truths of the sort p&q. This clashes with the Decomposability Principle.

Notably, the fact that the Decomposability Principle clashes with the modalist analysis stems from a deep feature of modalism. The idea is *not* that (Modalist) analyses non-accidentality in terms of the holding of a long conjunction. This would be a slightly off way of understanding the charge, because technically (Modalist) understands non-accidentality in terms of an (generic) quantification over an infinite number of worlds. The problem is with the nature of the modal stuff quantified over – the possible worlds in question. Irrespective of how we conceive of these entities (whether they are sets of propositions, concrete worlds…), possible worlds semantics treats them as discrete entities, such that no relations (besides similarity relations) may hold between them. Consequently, facts about accidentality are treated as ultimately decomposable into facts about truths at separate independent worlds.

It does not help either to require that there be some connection between p and q at a world, as most modalist theories can (and do) admit. For example, a popular of phrasing the modalist idea is by saying that p is non-accidentally connected to q if p causes q and all the relevant p-worlds are worlds in which q obtains *on the same basis* (namely p).[152] For it will still hold in this case that there might be no discernible relation between the truths at each world – that the pattern of p causing q across a modal subspace holds no meaningful information as to how p and q are actually related.

Let's see this at work for reasons-responsiveness. For the fact that the agent φ-ed (A1) and her φ-ing matched with her reason to φ (A2), the modalist offers the following analysis: It is non-accidental that [A1&A2] iff a sufficient proportion of the relevant worlds in which the agent φ-s are worlds in which her φ is caused by R (thus, a sufficient proportion of A1-worlds are A2-worlds). The idea is that the fact that the agent φ-s and the fact R causes their φ-ing are companions (and thus, so are A1 and A2). Thus, the fact

[152] See for example Pritchard (2007,2008); Sosa (2007), 26.

that [A1&A2] decomposes as follows:

[A1&A2]	(i) R causes A1
	(ii) R causes A1 in w1
	(iii) R causes A1 in w2
	.
	.
	(…) R causes A1 in wn

This cannot be what non-accidentality consists in. Facts about non-accidental [p&q) cannot decompose into separate non-relational facts p and q.

Moreover, importantly this structure opens modalism up to the corresponding exploits, the exploits that establish the relevant truths at each world without establishing any relation between these truths. These exploits are the kind of examples I discussed in chapters 2 and 3, examples of finking, masking, and mimicking, to which most importantly Frankfurt-like cases belong. Thus, my argument has additional explanatory power. It explains why all orthonomy notions are plagued by the same examples, and why no way of specifying what exactly the relevant worlds are seems to get rid of them.

Let me explain.

Throughout this chapter, I have referred to the central elements of cases of coincidence as exploits. Exploits always exploit the possibility of understanding an orthonomy notion compositely on some level by making the component facts or explanation independently true. This insight allows us to classify a wide range of cases as interestingly connected.

In cases of *mere coincidence* the fact that [p&q] is decomposable into two separate independent facts p and q. *Simple exploits* exploit exploit this decomposability by telling a story in which it is true that p and it is true that q without there being any further relation between p and q. They thereby undermine attempts to analyse a non-accidental relationship between p and q in terms of a simple composite account. The story of the person who believes that p because they were hit in the head by shingle while p is also true is such a story.

In cases of accidentality, the fact that [p&q] is decomposable into two separate independent more specific facts [p1&q1] and [p2&q2] (depending on description). *Advanced exploits* exploit this decomposability by telling a story in which it is true that [p1&q1] and it is true that [p2&q2] without there being any further relation between

these two facts. They thereby undermine attempts to analyse a non-accidental relationship in terms of a description-unspecific relation such as causation. The stories told in cases of deviance are of this sort. In them the agent φ-s and their φ-ing matches with their reason in one sense (a causal sense) and the agent φ-s and their φ-ing matches with a reason in another sense (a rationalising sense). But there is no connection between the two senses.

Finally, in cases of (modal) accidentality, the fact that [p&q] is decomposable into a range of separate independent facts of the form 'p causes q' across a relevant modal subspace. Modal exploits exploit this decomposability by telling a story in which each truth at a world is turned on (or off!) without there being the appropriate connection between these truths. These stories involve devices, guardian angels or wizards – elements which establish the relevant truths at each world without relating them in the appropriate way. Examples of finking, masking, and mimicking belong to this class – as do Frankfurt-cases.

The way in which I presented these categorisations perhaps makes it seem that the facts about accidentality decompose in *either* of these ways, when of course they may decompose in *all* of them at once. Which way of decomposition is important is determined by what exactly we are asking about in our coincidence questions. Orthonomy notions are threatened by all three forms of decomposability, because we are evidently asking highly specific coincidence questions about them. The similarities of these exploit cases however also mean that we can easily translate between them, which exposes some interesting relationships of cases across subdisciplines. Let me just demonstrate this at the example of the problem of deviance.

So-called sensitivity solutions to the problem of deviance, developed mostly for the notions of 'action' or 'acting for reasons' (Morton 1975; Peacocke 1979a; 1997b[153], 128 ff; Shope 1992, 256 ff) start from the assumption that what is lacking in deviant cases is that the agent's action is sensitive to reasons. Sensitivity of course is most straightforwardly spelled out in counterfactual terms. For the action to be sensitive to the reason is for it to be true that if the reason had been slightly different, the corresponding action would have been slightly different as well. Or, as Bishop puts it:

> "[…] over a sufficiently wide range of differences, had the agent's intention differed in content, the resulting behaviour would have differed correspondingly." (Bishop 1989, 150)

[153] Peacocke is often discussed as a modal sensitivity account, although his own examples (discussed below) and explicit statements distinguish him from modal accounts. His account is that a differential explanation must obtain between R-belief and action. Differential explanations and modal profiles can come apart.

We already know this approach. It is, with slight tweaks, a version of the reasons-responsiveness models I criticised in chapter 2. To be sensitive, or responsive, it maintains, is just for a range of counterfactuals (or similar modal truths) to hold.

However, these approaches are unsurprisingly afflicted with the same type of cases that I have discussed exhaustively in chapter 3. They are afflicted by cases which show the modal profile of a φ-ing is irrelevant to whether it counts as deviant or non-deviant:

Christopher Peacocke, in defending the advantages of his "differential explanation" version of the sensitivity proposal, presents the following case:

> Suppose that a person perceives an array of objects in a television studio which is insulated from all outdoor light. We can suppose too that these objects affect light-sensitive devices which are connected to the studio lighting in a way that if any one of them perceptibly moved or visually altered, then all the lights would go out and nothing would be perceived. Hence it would not be true that if any one of the perceived objects altered, there would be corresponding change in the person's experiences. But this does not in any way throw doubt on the claim that the person in the studio perceives the array of objects: it is a clear case of perception. (Peacocke 1979a, 76)

This is a kind of example, in the domain of visual experience, where a causal chain is non-deviant despite the relevant counterfactuals being false. If the example looks familiar, it is because it has the same structure as Frankfurt-style examples as well as my blind spot cases extensively discussed in chapter 2.[154] The agent is actually successful in their endeavour in these cases – they actually perceive, or act for reasons, depending on the domain in which the case arises – but the relevant counterfactuals are false (see Mantel 2018, 33 for the same type of case against counterfactual sensitivity).

The reverse cases are also present in abundance. Take for example the following case from Mantel (2018), which is a variation from a deviant case in which Bob has a reason to scream (because this is the only way to save his life), which makes him so angry that he screams:

> Guardian Angel
>
> Suppose Bob's motivation to scream is caused by his anger, just as it was in the original example Angry Bob. However, Bob's guardian angel is right next to him

[154] Bishop (1989), 151 also discusses general Frankfurt-type structures in relation to sensitivity solutions to deviance. Stout (2010), 163 even turns the dialectical path of this thesis around, suggesting that we can use the Fischer and Ravizza model of reasons-responsiveness to solve causal deviance. But clearly, given that causal deviance is an actual-sequence phenomenon , and Fischer and Ravizza's model cannot properly handle such phenomena, this suggestions is not promising

> and if Bob had not been motivated to do whatever he believes is the only way to save his life, his guardian angel would have inserted in him the motivation to do just that. If Bob had not believed that screaming was the only way to save his life, but instead that writing the word 'HELP' into the sand was, then his guardian angel would have endowed him with the motivation to write the word 'HELP' into the sand, and likewise for any other belief about how to save his life. However, the guardian angel does not actually interfere since Bob arrives at the right motivation due to the causal chain involving anger. (Mantel 2018, 30)

Here, we have case in which the relevant counterfactuals are true, but the actual sequence exhibits a deviant causal chain, nonetheless. Again, the example should look familiar. It is the example of a dispositional mimic, a device (or in this case angelic being) that feigns an objects possession of a disposition by making the relevant counterfactuals true. Only in this case, what the device mimics is not the *possession* of the disposition, but its manifestation (so masking and mimicking come down to the same thing for actual manifestations). It is the same kind of case, moreover, that we encountered while assessing the suggestion that all we require is an explanatory connection between the two roles a reason plays in explaining action.

So since deviance is an accidentality phenomenon, it exhibits the same recalcitrance to being analysed modally – either in counterfactuals or in possible world terms. Masking cases sometimes obscure the fact that non-accidentality exhibits an aversion to modal analysis, because they feature circumstances in which an agent merely *can* – has the ability to – do something despite the relevant modal profile not being true of them or vice versa. Cases of causal deviance have a better chance to prompt philosophers to think about non-accidentality as an actual-sequence phenomenon. For they feature circumstances in which the agent actually *does* something, without the relevant modal profile being true of them or vice versa. Evidently, the two phenomena – masking of abilities and deviant explanatory chains – fall within the same case-class. Both involve exploits, both ultimately concern intuitions about accidentality.

Recall that in masking cases the object retains its disposition, the typical trigger conditions obtain, but the object does not manifest its disposition. All of these criteria apply in deviant causal chain cases. For even if they feature a causal outcome that matches the typical manifestation type of the disposition involved, their very point is that this outcome does not amount to a manifestation of the relevant disposition. So deviant causal chain cases are a subclass of masking cases. A curious detail about deviance cases is they feature causal outcomes, while masking is normally seen as involving the *lack* of causal outcome. The typical example of a fragile glass wrapped in bubble wrap involves the glass *not* breaking. But deviant causal chain cases show us that this feature of typical

masking cases is relatively inessential to their lesson. Lack of causal outcome is just one way in which a disposition may fail to manifest. Another way is for the disposition to be overridden by another causal process – or indeed another disposition with the same manifestation type – which produces a causal outcome that just happens to fit the manifestation type of the pre-empted disposition.[155]

To summarise: The compositeness argument against modalism explains why modalism fails. It also explains why Frankfurt-type counterexamples keep appearing in all of the epicycles of modalist analyses (as seen in Chapter 3). Modalism is deeply structurally incapable of getting rid of cases of accidentality. Thus, *no* version of the view, no matter how sophisticated its modal analysis, will be able to get rid of modal exploits – and therefore Frankfurt-type cases. For modal exploits exploit the very fact that modalist solutions are composite. And modalism cannot get rid of compositeness, for compositeness follows immediately from the syntactical properties of possible worlds (no matter how we understand their metaphysics). Possible worlds are discrete entities. But this means any arrangement of p and q across a subspace of worlds, no matter how significant looking, will not make available the relational information we are seeking to express when we find that p and q are non-accidentally connected. Facts about non-accidental connections are not about the pattern of two separate items, they are about the synthesis or fusing or of these items. They are not about companions; they are about soulmates.[156]

This is why modalism about non-accidentality cannot succeed, and why it is irreconcilable with the explanationist account. Modalism represents non-accidentality as a composite. Explanationism does not.

With this, I have kept the promise of chapter 3 to develop an account of the non-accidental relationship between reason and action. However, it is not obvious, from what I have said so far, how the contents of this chapter connect up to the contents of the last chapter. That is, it is not clear how my treatment of responding to reasons as the exercise of the capacity to respond to reasons links up with my explanationist account of non-accidentality. The next chapter will provide the necessary bridging information by developing an account of what it is to exercise a capacity in explanationist terms.

[155] One further consequence of this insight is that deviance is a relative notion. Depending on which dispositional system we focus on, a process may count as both deviant and non-deviant. The shattering of the glass in Molnar's example may count as non-deviant of it is understood as the manifestations of the larger dispositional system which contains both the Z-ray emitter and the glass.

[156] Again, this metaphor comes from Felipe and Tognazzini (2010)

Chapter 7:

Exercising Capacities and Non-Accidentality

Part I
Exercising Capacities

1. Introduction

The previous chapter introduced a general theory about what it is for two facts to be non-accidentally connected, including when these facts are about the agent's reasons and actions. Thus, what the previous chapter has given us is an account of the non-accidentality essential to reasons-responsiveness. If my account is successful, then it provides an understanding of the essential non-accidentality in responding to reasons that does not rely in any way on alternative possibilities. It instead illuminates non-accidental relationships in terms of explanatory properties alone. But this leaves, at the very least, a gap in my presentation. For in chapters 4 and 5, I presented an account of responding to reasons based on the notion of an exercise. I did say in those chapters that it seems to be part of the concept of an exercise that it picks out a non-accidental connection. However, since I left the concept unanalysed, I did not say anything about *how* it automatically picks out non-accidental relationships. With my account of non-accidentality on the table, I can fill this gap, which is the purpose of this chapter. The chapter presents a structural (explanationist) account of what it is to exercise a capacity and shows how the account naturally integrates with my understanding of non-accidentality. I thereby hope to shed more light on the nature of manifestations/exercises and their connection to non-accidentality.

The chapter is split into two parts. The first, bigger part introduces my explanationist account of exercising. The second, smaller part contains an overview over other views on exercising dispositional properties. I will especially contrast two opposing approaches to how the exercise of dispositional properties integrates with a broader metaphysical picture. Focussing on these approaches will help contextualise the project of this thesis by situating it in a larger dialectics, and it will help address nagging questions about how the explanatory properties that I have used in this thesis to

illuminate the notion of responding to reasons relate to metaphysical, in particular causal matters.

Let us now start with part one, which addresses the link between exercising capacities, explanation, and non-accidentality.

2. Dispositions in Explanation

Glasses break and matches light. Perhaps more importantly, agents respond to reasons. What all of these cases have in common is that they involve the manifestation of a disposition.[157] The overarching question of this chapter will be what manifestations are.[158] That is, I will try to answer the question what it is for a φ-ing to be the manifestation of the disposition to φ. This question inquires about the link, as it were, between a disposition and its manifestation. I will be convenient to call this link the *manifestation relation*.

How should this question be approached? I think it will be helpful to start with those features of manifestation concepts that we have the most immediate grasp on. We are familiar with how breakings relate to fragility and lightings relate to flammability on the basis of how we use those concepts in our explanatory practices. That is, sometimes, it seems, we can understand and make understandable the breaking of a glass in virtue of the fact that the glass is fragile (see Dretske 1988, Kim 1988, McKitrick 2005). And there is certainly nothing strange or artificial about sentences like: 'the glass broke because it was fragile', 'the match lit because it was flammable' or even 'the student stayed silent in class because they are shy' or 'Razvan chuckled because he was nervous'.[159]

Indeed, that the fragility of the glass affords us some understanding of why the glass broke seems to be a common sense default to such a high degree that exceptional cases in which the relevant explanations are *not* forthcoming would appear like remarkable flukes to us. We rely so much on the explanatory role of fragility that we often wonder what could have possibly make the glass *not* break in those rare instances when we

[157] Some of the topics in this chapter will be related to how we spell out these properties – whether, for example, we think that they reduce to collections of counterfactual conditionals or not. These issues will become important in the second part of the chapter only however, which is why I will be referring to dispositional properties with no specific analysis of them in mind.

[158] This question is easily misunderstood as a question about the ontological category of manifestations. I am not asking about that. I am asking about the nature of the relation between manifestation and disposition – irrespective of what kind of thing manifestations are. Answers to the former question include that manifestations are effects (McKitrick 2010) and contributions to effects (Molnar, 2003: 195, Mumford, 2009: 104).

[159] Explanations in terms of character traits are often treated as dispositional explanations, especially by those who think desires and beliefs are dispositions (see for example Martin 2007). For a treatment of character trait explanations similar to the account of the manifestation relation I advance below, see Morton (1980), Fileva (2016). See Alvarez (2017) for scepticism about the thesis that character traits are dispositions.

smash it to the floor and it bounces. Whatever our eventual account of the link between disposition and manifestation turns out to be, it should preserve this minimal insight about dispositions in explanation:

> **Explanatory Link**: If some φ-ing is the manifestation of a disposition to φ, then the disposition to φ in part explains the φ-ing.

But even this minimal guiding principle for our inquiry might be considered controversial. At least two initial obstacles could be raised against the idea that dispositions have a role to play in some explanations.

First, dispositions have famously raised the suspicion of being explanatorily void. Perhaps most famous is Molière's (1935) mockery of the *virtus dormitiva* of opium, in which the question why opium makes one sleep is answered by pointing out that it has the power to cause sleep. The nature of this charge is clear: When we explain why opium makes someone sleepy by the fact that it possesses the power to make people sleepy, we have explained nothing.

But we need to heed our wording here. Explaining why some object possesses a certain causal power is not the same as explaining why some typical manifestation occurred. My interest here is in the latter question, and I agree that the former question has metaphysical import and remains unanswered by merely pointing back to the power. However, perhaps Moliere was being sloppy. Perhaps his real complaint was that explaining why someone fell asleep by appealing to the power to make someone fall asleep is explanatorily void. This is certainly how the complaint is interpreted nowadays anyway, with little regard to the original formulation (Kistler 2007, see Michon 2007 for an exception).[160]

Why is it no explanation at all to say someone fell asleep because opium, when ingested, leads to sleep or that the glass broke because it was fragile? The reasoning must be that 'fragile' just means 'liable to breaking in the right circumstances', so to say that the glass broke in the right circumstances because it was fragile is to say that it broke in the circumstances because it was liable to break in the circumstances. Such an iteration of the manifestation statement we are seeking to explain certainly looks circular, and circularity does not afford the understanding needed for an adequate explanation.

But it looks like this reasoning goes through only if we confuse the dispositional property of fragility with the name of this property (Kistler 2007). 'Liable to break in the right circumstances' is a name we can give the dispositional property of fragility. We can thus

[160] Similar arguments to this circularity objection advanced amongst others, by Block (1990); Dardis (1993), and Jackson (1995), target the causal relevance of dispositions based on their semantic features.

create vacuous looking statement by replacing instances of predicates for that property with 'liable to break in the right circumstances'. But this does not establish that the property itself does not have a role to play in bringing about and thereby explaining the relevant manifestation. As Kistler puts it:

> Opium is not a placebo. The problem stems from the functional meaning of the expression 'dormitive virtue'. It determines the identity of the property only indirectly, by its typical effect. This does nothing to put in doubt its reality or causal efficacy. (Kistler 2007, 128)

There is a remaining worry here though, that we should take on board. The rebuttal I just gave depends on the assumption that in the right triggering circumstances, the fact that the relevant object has a disposition, like fragility, will play some role in explaining why, *given that the right circumstances obtain*, the breaking occurs. But of course this means that we should interpret the sentences I started with, sentences like 'the glass broke because it was fragile' as elliptical. What this sentence really says, with its ellipsis spelled out, is: 'the glass broke because it hit a hard object and it was fragile'.[161]

This concession leads directly into the second point that might make us doubt the explanatory efficacy of dispositions. For it seems like for every explanatory sentence that states a dispositional condition as its explanans, we can find a perfectly adequate replacement sentence that *only* mentions the salient triggering circumstances. For the fragility of the glass, we can say 'the glass broke because it hit a hard object'. Similarly, we often say 'the match lit because it was struck' and 'Razvan threw his computer out of the window because it vexed him'. These sentences don't explicitly mention dispositional properties. Yet they seem to offer a perfectly adequate explanation of the target phenomenon, i.e. the manifestation. The worry here is not that dispositions can't make an explanatory contribution by definition, but that there is no explanatory work left for them to do.[162] After we have provided the triggering circumstances for the disposition, the thought goes, mentioning the dispositional property itself will add no new relevant information and so will not increase understanding of the target phenomenon. Here, the charge isn't that dispositions are explanatorily void, but that they are explanatorily idle.

There is another way of understanding this worry that will prove illuminating: the worry is that if dispositional explanations are elliptical for statements like 'the glass broke

[161] Maybe the disposition acts as an enabling condition for the causal relation? I discuss a proposal like this in section 9.

[162] This argument is closely related to what is known as the Exclusion argument, put forth most famously by Prior, Pargetter, and Jackson (1982), but which also appears in Kim (1990)

because it hit a hard object and it was fragile', we are facing an overdeterminiation problem. The fact that the glass broke cannot be explained both by the dispositional property and the triggering condition, given that at least the triggering condition alone seems to provide a perfectly adequate explanation (see McKitrick 2005, especially section 8 for a discussion of the role of overdetermination).

The problem cannot be resolved, moreover, by simply holding that the trigger and the disposition explain the manifestation *together*. For it seems what is presenting us with the key problem is exactly that we have no account of what contribution exactly the dispositional property makes in this collaboration. What we are looking for, in other words, is a *job description* for the dispositional property in explanations.

Here is a way to make this result visible: initial reflection upon the two reservations about dispositions in explanations leads us to the view that dispositional explanations have the general form: *manifestation because disposition and trigger*. Later, I will argue that this in fact *not* the form we should assume. But for now, it is a convenient vehicle to express the most pressing issue with dispositional explanations: the current form does not make visible which explanatory role the disposition plays.

Hence, the two initial worries just presented don't so much undermine (Explanatory Link) as they refine it. They suggest that what should guide our inquiry into the manifestation relation is not only the minimal insight that manifestations are partly explained by dispositions, but also the explicit desideratum that whatever account of the manifestation relation we end up with, it must provide us with an understanding of *how* dispositions explain, or what specific explanatory role they play:

> **Job Description Requirement**:
> Any account of what it is for some φ-ing to be the manifestation of the disposition to φ must spell out what job the disposition to φ has in explaining the φ-ing.

This requirement will guide my development of an account of the manifestation- relation in the following. But it will not be the only requirement doing guiding work. The other important guideline will be provided by how the notion of a manifestation interacts with cases of deviance (and thereby accidentality). In order to set up this guideline, I will look at the connection between the notion of a manifestation and the problem of deviance now.

3. Deviance and Manifestation

There has recently been a large-scale convergence across many fields in philosophy towards the idea that the problem of causal deviance can be solved with recourse to the

notion of the manifestation of a capacity (Arpaly 2006, 46-48; Hyman 2015, ch. 5; Lord 2013, ch. 4; Mantel 2018, ch. 2 and 8; Marcus 2012; Mayr 2011, ch. 5; Millar 2019; Setiya 2007,23; Schlosser 2011; Smith 2009; Sosa 2017; Stoecker 2003, 313; Stout 1996, ch.3; Stout 2005, ch.6; Stout 2010; Turri 2011, 390-393; Wedgwood 2006, 664–667).

The basic idea common to all of these accounts is that the special kind of explanation they require in their analysis of a given notion is what I labelled (ch. 5) an *exercise-explanation* or *manifestation-explanation*, i.e. those explanations in which we explain some occurrence by reference to the fact that it was a manifestation or exercise of a disposition. In fact, my own account has in part relied on the insight that exercising capacities provides the right type of explanation. However, in the last chapter I developed a structural account of non-accidentality, which provides an analysis of the problem of deviance that does not require recourse to the notion of an exercise. In this section, I want to suggest that accounts of the manifestation-relation should be able to explain how the notion of an exercise eliminates the possibility of accidentality – in my terms: it should be able to be integrated into my structural account of non-accidentality.

The core idea of manifestation accounts is that the notion of an item being the manifestation of a capacity or disposition already carries with it implicitly a criterion for picking out the right sort of explanatory link between, say, reason and action. That is, for some φ-ing to be the manifestation, rather than a mere symptom (Hyman 2015, 117) of the disposition to φ, it must already be linked explanatorily to the disposition in the correct way. Dispositions are also said to be "process-specific" (Molnar 2003, 91), that is, only a specific kind of causal process can count as the exercise of a given disposition.

The idea of manifestation accounts is then that in, for example, Davidson's climber case, the connection between mental state and causal outcome is deviant because the climber's letting go does not manifest the capacity for intentional action. Only etiologies that feature manifestation of the right type of dispositional property are eligible for the relevant notion in question.

However, I pointed out in the last chapter that all orthonomy notions face deviance problems. Since the notion of an exercise itself shares essential features with orthonomy notions, a curious dialectical situation for the manifestation solution to deviance is created. For plainly, if the notion of a manifestation is *itself* an orthonomy notion, it will itself be subject to deviant causal chain cases.[163] Here is one such case presented by Molnar:

[163] Lewis (1997) tries to argue that dispositions do not exhibit the problematic process-specificity but see Molnar (1999) and Molnar (2003) for strong responses to Lewis argument.

> Suppose that knocking some object causes 'Z-rays' to be beamed on it, which in turn causes it to shatter in the way fragile things shatter when knocked. Here an s–r sequence occurs that satisfies the analysans, yet the shattering is not a manifestation of the fragility but of a deviant process. (Molnar 2003, 91)

Perhaps Molnar's case is too sci-fi for you to elicit stable intuitions. Nonetheless, it should be clear that the case gives us a blueprint for constructing less outrageous scenarios. All we need is the insight that just like reasons or belief-desire pairs, dispositional systems can cause and thereby explain things in more than one way.

Consider for example the small machine designed to open your garage door. We may conceptualise this machine as the dispositional system the manifestation of which is the opening of the garage door – it is a disposition possessed by the door to open when triggered. But now consider that you trigger the machine remotely in order to open the garage door, but an internal failure makes the machine produce a strange loud rustling sound. Alarmed by this sound, you manually open the door and check what is going on. Here, the typical dispositional trigger of the garage door's disposition to open causally explains the opening of the garage door. But the opening of the garage door is not a manifestation of the door's disposition to open. We can imagine all sorts of 'misappropriations' of dispositions like this. Not all of them need to involve intermediary agents and not all of them need to involve intermediary steps[164], as the literature on deviance has proven (Mayr 2011, 114 – 117; Bishop 1989).[165] And just like for all the other orthonomy notions, the deviance problem for dispositions "[…] draws attention to the fact that dispositional dependence contains something over and above causal dependence." (Molnar 2003, 91)

Given that the link between disposition and manifestation it itself liable to infection by cases of deviance, the notion that we can solve deviance by reference to the concept of a manifestation must at least seem strange. At the very least it isn't very informative to say that the right kinds of explanations are those which involve the manifestation of the relevant dispositions when what we have to say about the manifestation of a disposition is that it is grounded in the right kind of explanation that connects disposition and typical outcome of that disposition. According to some versions of primitivism about the manifestation relation (see section 8) this move might not be straightforwardly circular because the idea is that exercise-manifestations implicitly *pick out* the right lower-level connection between disposition an outcome. But it is still not very informative. We should aspire to, and I think to some degree we can, do better. Minimally, I think we

[164] Smith (1977) and Prior (1985) attempt to get rid of dispositional deviance by requiring 'direct' connections.

[165] Brand (1984), and Mele (1992a) develop anti-deviance strategies based on causal immediacy.

should offer an explanation as to *why* exercise-explanations have the curious feature of preselecting the right, non-deviant etiologies. It is precisely this task that I will tackle in the next sections. In fact, I hope to show that explaining why the notion of manifestation comes with an inbuilt exclusion of deviance will provide us with an account of the manifestation relation that satisfies the job description requirement in a comprehensive way, i.e. that points out precisely what explanatory role dispositions play. Still, it is worth formulating the ambition as its own guideline:

> **Non-Accidentality Requirement**:
> Any account of what it is for some φ-ing to be the manifestation of o's disposition d to φ in C must explain why φ-ing being a manifestation of d makes it the case that it is not a coincidence that o φ-ed in C and o had the disposition to φ in C.

In the next section, I will develop an account of the manifestation relation that is guided by the Job Description Requirement and the Non-Accidentality Requirement.

4. The Doubly Explanatory Account

If spelling out the manifestation relation is itself subject to the problem of deviance, then the structural approach to deviance I developed in the last chapter should be able to shed some light. Recall that I understand accidentality in terms of unsuccessful answers to coincidence questions. The first thing we need to figure out for the deviance cases about the manifestation-relation is then what the relevant coincidence question in the background is.

In order to do that, look at the deviance cases above again. When the glass breaks because the rays are beamed on it, in what sense does this event involve a coincidence?

Here is a redescription of the case I find illuminating: It is a coincidence, in the case of the glass, that the typical[166] triggering circumstances for fragile things to break cause the behaviour the glass is intrinsically disposed to exhibit (.i.e breaking). Thus, the coincidence question implicit in the scenario enquires about the relationship between the fact that C causes o to φ and the fact that o has d to φ in C, i.e. that C is the typical trigger for a disposition that o possesses.

The situation is structurally identical to Davidson's climber case reviewed in the last chapter. In that case, the climber's intention causes and rationalises their behaviour, but

[166] As I discussed in chapter 2, typical trigger conditions for dispositions and capacities should be given in terms of test-cases. Test-cases are alternative possibilities which serve as laboratory conditions for the relevant disposition – circumstances purged of interferences. I also discussed how interferences cannot fully be purged from test-cases.

it plays those two roles independently. In the deviance cases for manifestation, C causes o's φ-ing and C is also the typical trigger for o's disposition to φ in C. But it plays those roles independently. For C causes o to φ via a deviant route, and C is the trigger for o's disposition to φ in C independently of its causal role. Consider the glass again. The hitting of a hard object causes the glass to break (because it causes the beams to be shot at it). The hitting of a hard object is also the typical trigger condition for its disposition of fragility. But the causal role that C plays with respect to the breaking of the glass is independent from its role as a typical trigger condition. This is why, I submit, we have the impression that the breaking of the glass is not a manifestation of its fragility. It is coincidental that the glass does what it is disposed to do in circumstances in which it typically does it, because it is coincidental that what causes the breaking is also what typically triggers the disposition to manifest (as a breaking).

So far so good. In the previous chapter, I also developed a view on what is missing from cases of accidentality. What is missing, according to the explanationist account of non-accidentality, is the availability of a maximally unified explanation. An explanation 'p because q' is maximally unified, iff it cannot be decomposed into separate independent explanations in terms of separate independent roles played by p and q. Thus, the titular unification refers to the existence of explanatory connections between all descriptions i of p and q such that q plays no separate explanatory role under any i. Now we only need to apply this principle to the manifestation deviance cases. The intuitions here already point in the direction of my proposed structural solution. Intuitively, the breaking of the glass in Molnar's case does not count as the manifestation of fragility because hitting a hard object does not cause the breaking *as* the trigger of the dispositional system but fulfils its causal role separately. Thus, what is missing from the case, intuitively, is that hitting a hard object causes the glass to break *because* it is the trigger of the disposition of fragility.

There is a more illuminating way to put this last insight, however. What the intuition – and my account – are aiming to express is that when some φ-ing is the manifestation of o's d disposition to φ in C, then d explains why C caused o to φ. This is after all what we mean when we say that C must cause o to φ in its capacity as o's trigger. What we mean to express is that C's causal influence on o's φ-ing must reflect the involvement of o's dispositional properties.

This insight has surprising consequences for the Job Description Requirement. For it turns out that I have just given an intelligible job description to the dispositional property. It isn't the case, in this picture, that the dispositional property *and* the trigger circumstances explain the object's φ-ing (as I had initially assumed in section 2). Rather, the dispositional property explains why C caused o to φ.

Importantly, when I say that d explains why C caused o to φ, this explanatory connection is to be read as picking out a unified explanation of why C causes o to φ, in accordance with my account of non-accidentality. This is important in order to exclude cases, reviewed in the last chapter, in which accidentality returns because even though d explains why C caused o to φ, it explains o's φ-ing based on C in the wrong way. What needs to be secured is that there is *no* explanatory role of C in which C causes o to φ independently of its role as a dispositional trigger. For this possibility will immediately open the gates for cases in which it is ultimately accidental that o φ-s in C and o has the disposition to φ in C. Hence, we need to specify that in a proper exercise explanation, d explains in a maximally unified way why C caused o to φ. We can summarise this as an informative principle about manifestation:

> **The Doubly Explanatory Account (imprecise):**
> Some φ-ing is the manifestation of o's disposition d to φ in C iff d explains$_{\text{maximally unified}}$ why C causes o to φ.

I call this account 'doubly explanatory' because C causing o to φ by itself of course already offers a causal explanation of o's φ-ing.[167] But as I discussed in the previous chapter, such an explanation will not be of the right kind, for it is forthcoming even in deviance cases. What transforms cases of deviance into cases of non-deviant connection is the additional dispositional superstructure in d's explaining why C caused o to φ.

It is worth emphasising that the Doubly Explanatory Account thereby introduces a subtle but important shift in explananda and explanantia, that isn't currently very well represented in the schema. Note that breaking and the other manifestation types discussed here can be expressed in transitive and intransitive verb forms. I.e. we can either say that C broke the glass, in which case it is entailed that C caused the glass to break. This is the transitive ('causative') verb form. Or we can say that the glass broke, which is the intransitive verb form.[168] When we use the verb in its transitive form, we usually express that an outcome or result occurred. When we use the verb in its intransitive form, we express, to put it neutrally, that the relevant activity is in progress. The Doubly Explanatory Account executes a shift between these two. In explanations of breakings of a glass in terms of the typical triggering circumstances for breaking that do not involve the fragility of the glass, all we do is explain a causal outcome. We explain why the glass broke in the outcome sense. According to the Doubly Explanatory

[167] The idea that we can explain why some factor explains another might be found mysterious. But I think it is perfectly ordinary. To explain why A explained B is to explain why A afforded us a new understanding of B. That is, we can explain the explanatory fact that A explained B by spelling out the reasons why A is an explanatory factor for B. However, if you feel uncomfortable with this, you can still admit that it is possible to explain why A *caused* B, which is in the official formulation of the view.

[168] The locus classicus for an extended discussion is Parsons (1990), especially ch. 6.

Account, exercise-explanations do not explain why the glass broke in the outcome sense. They explain why the glass was broken by the hard object in the process sense, i.e. they explain why the hitting of a hard object lead to the breaking of the glass.[169] Correspondingly, we need to distinguish between two ways in which we may refer to C. When the hitting of a hard object is conceived of as merely the cause of the breaking, then it causally explains the causal outcome (the breaking). It also explains, in a different independent explanation why the causal outcome matches the behaviour typically exhibited by the object in these circumstances. But in exercise-explanations, C is conceived of as playing both roles in a unified way. It is easy to misunderstand this as the claim that C now plays a unified role in explaining the same thing – the causal outcome. But the foregoing discussion entails that we should rather think of the difference between unified and non-unified explanations as a kind of explanatory ascent. When C explains the breaking in a unified way, it explains a relational fact, recall, and thus it does not explain just the causal outcome. It explains why C-causal caused that outcome. To distinguish these two ways in which C may show up in explanations, I'll refer to the unified role as 'C dispositionally bound' sometimes, because it captures the way in which is C is subsumed under a dispositional description in exercise-explanations.

The Doubly Explanatory Account can be restated more precisely with these clarifications:

> **The Doubly Explanatory Account (precise):**
> Some φ-ing (process) is the manifestation of o's disposition d to φ (process) in C iff d explains$_{\text{maximally unified}}$ why C-causal caused o to φ (outcome).[170]

These considerations make clear that the account does not only clarify the job description of the dispositional property, it also explains why the notion of a manifestation, as it were, automatically eradicates deviance. According to the Doubly Explanatory Account, dispositions are explanatory unifiers. To say that some φ-ing (process) is the manifestation of a disposition to φ in C is to say that C plays a unified explanatory with respect to the φ-ing (process). That is, to understand C (dispositionally bound) as 'the dispositional cause' of the φ-ing (process) is to pick out C under a very specific description, under which C cannot be decomposed into several independent explanatory roles which each explain aspects of the φ-ing (outcome) separately. Thus, to assume that C causing o to φ is a manifestation of o's disposition to φ in C is to pick out

[169] Thus, the account, especially when applied to the capacity to respond to reasons bears similarities to Dretskes dual explanandum approach to reasons-explanations (Dretske 1988, 1993, 2009).

[170] Some proposals in virtue epistemology are structurally very close to my own account. For example, John Greco proposes that in order for S's belief that p to amount to knowledge, it must both be true that S believes that p on the basis of an intellectual ability and S believes p is true because S's belief is produced by the ability (Greco 2010, 71)

a maximally unified explanation of o's φ-ing (process) in terms of C. And so, any explanation 'p because q' that can legitimately be expressed as an exercise-explanation will be a maximally unified explanation of p in terms of q. Since as I established in the last chapter, the availability of a maximally unified explanation is the mark of non-accidental connections, any explanation 'p because q' that can be expressed as an exercise explanation will establish a non-accidental connection between p and q. This is how exercising capacities automatically disperses the impression of accidentality: the notion of an exercise is structured so as to pick out maximally unified explanations.

To solidify the account, let me now apply it to the garage door case discussed above. In the garage door case, I remotely trigger the garage door to open, which makes the motor give off a strange sound, which makes me open the garage door manually. If you are worried that the case is too easy because it does not involve the dispositional system to open the garage as the proximate cause of the opening of the door, imagine I crawl in through a vent, take the door opening mechanism and use it to manually pry open the door. Now the remote triggering causes the door to be opened and the proximate cause of the opening is the dispositional system that typically opens the door when remotely triggered – except it is clearly being misappropriated. The Doubly Explanatory Account holds that in this case, even though the remote trigger causes the door to open and the dispositional property is causally involved in opening it, *the fact that the door has the disposition to open if remotely triggered does not explain (in a unified way) why the remote trigger caused the door to open*. This is because in order for the fact that the door has the disposition to open if remotely triggered to explain the trigger causing the opening, we have to assume that the remote trigger plays its causal role because it plays its role as dispositional trigger – i.e. it is dispositionally bound. But since in the example, C plays its causal role independently of it being true that it is the trigger for the door's disposition, the opening of the door (by the trigger) does not count as a manifestation of the door to open if remotely triggered.

With the account on the table, let me now address two applications of it that are important for the context for this thesis, namely (i) wat the account looks like for the exercise of the capacity to respond to reasons and (ii) how the account deals with cases of error (which have played an important roles in chapters 4 and 5).

5. Reasons-responsiveness as the Exercise of the Capacity to Respond to Reasons

In chapter 5, my contention was that responding to reasons goes along with the availability of explanations of the relevant φ-ings (my examples included actions and beliefs, I'll stick to actions here) in terms of those reasons. And the crux of my account

was that these explanations are exercise explanations – explanations that indispensably involve the agent's RR capacity. In the previous chapter, I pointed out that the explanandum targeted in coincidence questions about reasons-responsiveness – the rational action, as it were (where this description designates an relational fact) – just is the explanandum of reasons-explanations of actions. This is an unsurprising result, too. I am tracking a package phenomenon here after all: To say S's φ-ing was a response to S's reason for φ-ing is to say that it was non-accidental that S's φ-ing matched with that reason. And to say that it was non-accidental that S's φ-ing matched with S's reason for φ-ing is to explain why S φ-ed and S's φ-ing matched in terms of S's reason for φ-ing in a unified way. Hence, we can understand the non-accidental relationship crucial to reasons-responsiveness in terms of the way in which the reason explains the action in instances of responding to reasons. The account developed in this chapter allows me to complete this picture. For it allows me to pinpoint the role dispositional properties play in the conceptual scheme just outlined. It allows me to say, in other words, how the fact that reasons-explanations of actions are exercise-explanations makes them accidentality dispersing and thereby enables us to understand non-accidental relationships better.

Let us start with a standard reasons-explanation of action. Let us say that Pei-Lung jumped out of the window because his hotel was on fire. That is, he jumped out of the window for the reason that his hotel was on fire.

(Pei) Pei-Lung jumped out of the window because his hotel was on fire.

Sentences like (Pei), according to my account, will have a much more complicated structure than their appearance suggests.[171] First of all, when we provide a reasons-explanation of action, we are not just providing an explanation for the fact that Pei-Lung jumped. Many explanations are capable of explaining this fact, including those that mention irrational psychological or even external causes. What makes (Pei) a *reasons*-explanation of action is that it explains Pei-Lung's *rational jumping* – it explains the relational fact that Pei-Lungs jumping matched with his reason for jumping (namely the fact that his hotel was on fire).

Next, look at the explanans. We know now from cases of deviance (which just are cases of accidentality) that the fact that the hotel was on fire can explain Pei-Lung's action in the wrong way. This happens when the fact that the hotel was on fire fails to explain the relational aspect hidden in the explanandum. That is, the fact that the hotel was on fire may explain why Pei-Lung jumped, and it may also separately explain why Pei-Lung's jumping matched with his reason. For example, it might be the case that (recognition of)

[171] Given the extensive trouble they have caused and are still causing for the philosophy of action, I don't think it is a wild assumption that reasons-explanations of action have a complicated deep structure and should not be taken at face value.

the fact that his hotel was on fire made Pei-Lung so nervous that he jumped. In this case, the same fact explains both the jumping and the reasons-matching, but it doesn't explain the rational action. And so the explanation fails.

What needs to be added to account for the successful reasons-explanation reported in (Pei)? In (Pei), the fact that the hotel was on fire plays a maximally unified role as explainer. All descriptions of the fact such that the fact would play a separate explanatory role when so described are themselves connected by unified explanations. And we know now what has to be true for the reason to play this role. It has to be true that the reason is part of an exercise-explanations – or in other words: it has to be true that Pei-Lungs jumping was an exercise of the capacity to respond to the reason that the hotel is on fire.

Now we can slot in my explanationist account of the manifestation-relation. As I argued in chapter 2, the freedom-relevant capacities should be taken to be very specific capacities, depending on context. Pei-Lung's relevant capacity, for example, seems to be the capacity to respond to jumping-from-this-hotel-window reasons. This is the capacity to respond to sufficient reasons for and against jumping from the window of his hotel room. In fact, I think the description should be even more specific so as to include only reason to do with fire hazards (for reason to leave the room that are of an entirely different nature don't seem relevant to Pei-Lung's situation). Let's then say Pei-Lung manifests his capacity to respond to hotel-room-jumping-fire reasons. If Pei-Lung's jumping is an exercise of this capacity, the following is true: the fact that Pei-Lung possesses the capacity to jump out of windows in circumstances in which fire-reasons for and against jumping present themselves explains why the fact that Pei-Lung's hotel was on fire caused Pei-Lung to jump. This is merely an application of the Doubly Explanatory Account, according to which o's φ-ing is a manifestation of o's disposition to φ in C iff d explains why C caused o to φ.

Recall my explication of how the possession of the dispositional property makes for a unified explanation. The crucial idea is that in ascribing the exercise of a disposition or capacity, we are conceiving of a certain fact or proposition under a highly specified description, namely as the trigger of the relevant dispositional system. As soon as we drop the assumption that the relevant fact or proposition causes an outcome in virtue of it being the trigger for the dispositional system, it will be untrue that the fact causing the outcome counts as an exercise of that dispositional property.

Now, my point about (Pei) is that is just an *instance* of an exercise explanation like this. In it, Pei-Lung's jumping counts as an exercise of his capacity to jump if reasons for and against jumping are present because the fact that his hotel is on fire is thereby described

as playing a causal role because it plays a role as the trigger for Pei-Lung's capacity. Thus, (Pei) is equivalent with:

(Pei – Exercise) The disposition to jump if hotel-room-jumping-fire are present explains$_{\text{maximally unified}}$ why the fact that the hotel was on fire caused Pei to jump.

There is an essential misunderstanding about this account that it is most pressing to address, since it leads into some of its interesting action-theoretic entailments. You might be confused how (Pei – Exercise) could ever be equivalent with a 'normal' reasons-explanation of action like (Pei). This is because it seems obvious that what counts as the reason in (Pei) – the reasons-proposition *that the hotel is on fire* – is not the explanans in (Pei-Exercise). In (Pei-Exercise), a disposition is doing the explaining. Moreover, the explanandum in (Pei), I said, is the rational action, whereas in (Pei-Exercise) the explanandum is the fact that the reasons-proposition caused Pei to jump.

It is essential to understanding my account to see where this thinking goes wrong. As in the glass case above, we need to be careful to distinguish between two ways of understanding explanandum and explanans. When Pei-Lung jumps, this can express the fact that there was a jumping event (whose agent is Pei-Lung). But it can also mean that Pei-Lung went through the process of jumping, i.e. it can refer to the fact that Pei-Lungs reason caused him to jump. Likewise, what we refer to as 'the reason' in reasons-explanations of actions needs more distinguishing. It might refer to the reasons-proposition – the fact or proposition which provides the reason. But as we know from deviance cases, the fact alone that this proposition is explanatorily involved does not guarantee a unified explanation. In unified explanations, the reasons-proposition is dispositionally bound. Thus, in these explanations, the fact that the reasons-proposition is the trigger for an RR capacity explains the fact that the reasons-proposition caused Pei-Lung to jump (i.e. the normative role explains the causal role of the proposition). It thereby explains the relational fact that Pei-Lung performed the rational action by pointing out that the reasons-proposition caused Pei-Lung to jump because it was the trigger for Pei-Lungs hotel-fire-jumping capacity. It explains, then, why the action has the property of matching the reason.

Again, the idea is that the move from a non-unified to a unified explanation involves a kind of explanatory ascent. It involves a move from a cause explaining an event to a dispositionally bound condition explaining why the cause caused the event.

Thus, the worry that (Pei) and (Pei-Exercise) are clearly not equivalent because they involve different explanantia and explananda only makes sense as long as we assume that reasons-explanations of actions explain causal outcome events and the explanans

in reasons-explanations of action is the reasons-proposition. [172] But I reject those assumptions.

Thus, (Pei) is equivalent with (Pei-Exercise). (Pei-Exercise) spells out the structure that is concealed by the surface grammar of (Pei). This does not change the fact that (Pei) and straightforward explanations are true. (Pei-Exercise) just gives an account of their mechanics.

Two interesting action-theoretic results are entailed. First, if we still hold that action explanations explain, well… *actions*, then we must hold that actions are not outcome events.[173] They are instead processes. In action explanations, in other words, we explain causings, not the outcomes of these causings. That the outcome events attached to these causings occur is consequently something that is *presupposed*[174] in action explanations, not something explained.[175] I welcome this result. It fits the view developed in this thesis and it is independently supported by a flurry of action-theoretic considerations.[176]

Second, it follows from this view that objective normative reasons – if by this term we mean the propositions which provide such reasons (not the reasons-relation) – are not identical with the explananda of action-explanations. Hence, there is a sense in which it

[172] I have mentioned in footnotes before that I started this thesis with realism about reasons and found a more appealing theory midway through. The idea that the reasons-proposition is not enough to make for a proper reasons-explanation is responsible for this shift. For it suggests that normativity is in part 'created' when we additionally assume that the reasons-proposition is dispositionally bound, i.e. when we assume that the reason the proposition caused the agent to φ is that the agent exercised their rational capacities in φ-ing.

[173] The view is so common that it has been called the 'standard story' (Velleman 1992a), but it is usually credited to Davidson (1980) – especially since it contains powerful arguments based on logico-semantic considerations for the view that actions are events.

[174] My view thereby avoids what Constantine Sandis (2006) calls The Conflating View of Reasons and the Conflating View of Action Explanation. The former view holds that the reasons for which we act are the reasons why our actions occur. The latter view holds that whatever explains an action explains why the actional event occurred.

[175] The Davidsonian view that actions are events actually has some (underdiscussed) explanatory problems as well, which are important to the topics of this thesis. This is because part of Davidson's argument is that action-sentences have an existentially generalised structure. But action explanations, in which such sentences often figure, this structure causes problem. What we explain when we explain actions is *not* why 'there was some event such that…' We want to know about particular agents (see Hornsby 2004) and their token actions (this argument had to be cut from the thesis due to word count restrictions).

[176] Pioneers of this view include Von Wright (1963) and Chisholm (1964). Alvarez and Hyman (1998) provide a systematic account of these considerations against thinking that actions are events that still strikes me as the strongest overall argument against the view.

is seemingly untrue that 'reasons explain actions', which is an assumption routinely made in metanormatvity.[177] This result too, should be welcomed.[178]

The emerging picture then is this. In reasons-explanations of action, we are asking about a relational fact – the fact that the φ-ing matched with the agent's reason to φ. The φ-ing we want to know about is the process of the reasons-proposition causing the agent to φ. Hence, we are asking why this causing matched with the reason for φ-ing. In cases of accidentality, we can't explain this relational fact. We can only explain the fact that the outcome event occurred (via the reasons-proposition as a causal factor) and the fact that the φ-ing matched with the reason (via the reasons-proposition as a justifying factor). What we are asking about, however, is the relation between these two facts. The rational φ-ing is the process in which the reasons-proposition causes the outcome event because it speaks in favour of the process of bringing about that event. We explain this relational matter if we can (correctly) mention that the agent's capacity explains why the reasons-proposition caused the agent to φ. This is because to explain why the reasons-proposition caused the agent to φ is to understand the reasons-proposition as playing the causal role because it counts as the trigger of the relevant dispositional system – i.e. the system which picks up the relevant kinds of reasons. Hence, to explain why the reasons-proposition caused the agent to φ is to say that the reasons-proposition caused the agent to φ under the specific normative guise the dispositional system is described as picking up on. Thus, we get the result that it is non-accidental that the φ-ing process matched the reason. It wasn't an accident that the reasons-proposition caused the agent to φ and also made it rational to φ because in being caused to φ by the reasons-proposition, the agent exercised his capacity to respond to reasons.

I now want to extend this picture just a bit more, addressing some issues that came up in chapter 4.

[177] Philosophers that this thesis can be attributed to include Williams (1979), Korsgaard (1986, 10), Parfit (1997, 2011, 37), Dancy (2000), Stoutland (2001, 96), Heuer (2004, 45), Lord (2008), Broome (2009, 88). Mantel (2014, 2018) calls this 'the identity thesis', although the relation assumed is not clearly that of strict identity.

[178] Cases of deviance can be taken to be a strong reason for thinking that the explananda of reasons-explanations are not identical to normative reasons. After all, the normative reason stays the same across deviant and non-deviant scenarios, i.e. the same proposition may be involved in both deviant and non-deviant explanations of the same fact. For an overview over arguments against the identity thesis, see Mantel (2018), ch.6-9.

6. Triggers and Counterfeit Triggers

I believe the foregoing discussion offers an additional insight into why exercising need not (and should not) be considered a success notion such that to exercise the capacity to φ is to exercise it successfully (see chapter 4 for a discussion).

In order to get this discussion started, imagine a simple dispositional system: A soda machine that dispenses soda bottles if 50p are inserted. We can tell the same sort of stories I have already told in this chapter about the soda machine. For example, inserting a 50p coin might lead to the machine shaking uncontrollably due to some minor malfunction which causes a soda bottle to be dispensed. This is a case of deviance. In it, inserting 50p causes the machine to dispense a soda bottle, but 50p being inserted doesn't count as the dispositional trigger. Thus, it causally explains the fact that the machine dispensed a soda bottles independently of the fact that the machine has the disposition to dispense soda bottles if 50p are inserted.

However, another feature is also obvious about the soda machine: dispositional systems are rarely fully discriminate when it comes to their triggering circumstances. This is because they usually react to structural features of their triggers. The soda machine is triggered by objects that have the same weight, shape, and perhaps texture as 50p coins. But not all objects with these properties *are* 50p coins. It is possible to manufacture an object with the same structural properties but made from another material and with different imprints (and perhaps with another purpose in mind). When we insert such an object into the soda machine, it will also dispense a soda bottle, and, as I argued in chapter 4, we should consider this as an exercise of the disposition to dispense soda bottles if 50p re inserted. This is because even though the system was, in a sense, duped, it did react to the properties it usually reacts to, and so it exhibited perfect functionality in dispensing the bottle. We can call triggering circumstances that lie within this range of discrimination of the system *counterfeit* triggers because they mimic the real thing the system is disposed to react to.

If it is true that the system counts as exercising the capacity to φ in C even if C is counterfeit, then there must be exercise-explanations available in terms of counterfeit C conditions. And it seems that there are. When I insert a counterfeit coin, then the insertion of the counterfeit coin causes the dispensing of a soda bottle. Moreover, importantly we ca explain why the insertion of the counterfeit coin caused the bottle to be dispensed in terms of the system's disposition to dispense bottles if (real) coins are inserted. This is because we are saying that the counterfeit is a structural replica of the conditions the system is disposed to track/react to.

Therefore, there can be exercise-explanations in terms of counterfeit triggers, and these triggers can stand in non-accidental relationships to outcomes.

I think the capacity to respond to reasons works in the same way. We can imagine counterfeits of normative situations which exhibit all the markers of the real thing and therefore trigger a response on the part of the agent. In these situations, an exercise-explanation in terms of what looked to the agent like a sufficient reason is still forthcoming. This proposition will of course be a *mis*representation of whatever reality underlies it, but as such it will still slot into the capacity to respond to reasons. I.e. I don't find it far-fetched to say that this proposition plays a causal role in bringing about the agent's φ-ing, and we can explain why it plays that causal role by referring to the fact that it serves as the trigger of the agent capacity to respond to (real) reasons.

If for some reason you are worried that false propositions do not cause, let the causes be whatever these propositions are about. If for some reason you are worried that false propositions don't explain, let me point out that the complete unified explanation involves both the false proposition and the relevant RR capacity. The complete exercise-explanation consists in the RR capacity explaining why counterfeit-R caused S to φ.[179]

In chapter 5 I pointed out that some might be sceptical about this approach because counterfeit reasons are not 'real' reasons. But the point here is that even if they are not real reasons, they are structurally similar to reasons in a way that allows an agent to stand in exactly the same relationship to them as to they would to real reasons. That is, we can explain our reactions to counterfeit reasons in exactly the same way – in terms of an exercise explanation (a unified, accidentality dispersing explanation). As far as my account of responding to reasons is concerned, this fact is more important than the metaphysical quarrel over what counts as a real reason.[180]

There is however a broader metaphysical issue that my explanationist account of the exercise relation brings up. The account is, as it were, a higher order account of manifesting dispositional properties, because it remains silent, for the most part, about whether there is any special metaphysics that underpins or backs unified explanations (and thereby exercises of dispositional properties). And while I lack the space in this thesis to fully develop such a metaphysics[181], I want to point out some general ways in

[179] I also claimed that there are exercise-explanations for responses to reasons that fail to be good responses. In these cases, we can explain an action in terms of how it deviates from what the reason recommended.

[180] As I said, I started this thesis with a realist background. But I am now much more convinced that we can turn the order of explanation around and use the notion of an exercise to understand the notion of a 'real' reason. Rationally treating R as a reason on this account might be enough for it to count as a real reason. But I lack the space here to develop these thoughts.

[181] I tried but went 20.000 words over the thesis word limit set by my university. The account exists, ask me about it.

which metaphysical questions intersect with the explanationist approach I developed here. The remaining sections are dedicated to this project.

Part II
Metaphysical Underpinnings

7. Two Pictures of an Underlying Metaphysics

So far, I have presented a picture of reasons-responsiveness according to which we can understand the notion of a *response* to R in terms of a non-accidental relationship between reason and action, and we can in turn understand this non-accidental relationship in terms of the special explanatory relationship between reason and action. Accidentality dispersing explanations are maximally unified. They explain a relational fact about the action and the action's matching with a reason for the agent. This means that they have to contain specific information not only about why the action occurred *and* why it was rational, but about the relationship between those facts – they explain the unitary fact that the rational action was performed. Exercise-explanations, according to my view, offer this type of information. If we say that in performing the action that was also the rational thing to do, an agent exercised their capacity to do what is rational, we are offering an explanation as to how the fact that the agent performed the action and the fact that it was the rational thing to do are related. They are related via the influence of the dispositional property.

In this chapter, I have given what can be called a higher-order account of the influence of the dispositional property. Explanations in which dispositional properties feature provide the relevant relational unifying information because in them we need to conceive of C as the relevant causal inciting feature *qua* dispositional trigger of the case. Only when C plays both roles in a unified manner will the dispositional property play the appropriate role in the explanation.

This account is 'higher-order' because it leaves it open how explanatory matters relate back to the worldly matters they are presumably about. I am here assuming a general view of explanation such that explanations are representational vehicles, paradigmatically expressed as sentences of the form 'p because q'.[182] And if they are representational vehicles, then there is something they represent. This thought is most helpfully representable in terms of Strawson's distinction between causation and causal explanation. For Strawson (1985), the main difference between causation and causal

[182] These issues closely align with the distinction, initially made by Salmon, between 'ontic' and 'epistemic' conceptions of explanation. Epistemic conceptions (Bechtel and Abrahamsen 2005; Wright and Bechtel 2007; Wright 2012) hold that explanations have an epistemic purpose – they make things understandable (or predictable, or intelligible). Sometimes explanations are even cast as a pragmatic phenomenon (e.g. Faye 1999). Ontic conceptions (Craver 2007) hold that explanations are the worldly relations that underpin our explanatory representations. I chose not to engage in this debate because it seems ill-conceived to me.

explanation lies in the fact that the latter is an "intellectual or rational or intensional relation" which "holds between facts and truths" (115) while causation is a natural extensional relation which holds between things we can assign "places and times in nature" (ibid). Things that can be assigned 'places and times in nature' are presumably particulars – datable, unrepeatable instances.

Strawson's distinction does not specify the relation between causation and causal explanation, however. For instance, given that we have accepted the initial distinction, we might ask how causal sentences like "The lighting strike caused the power outage" are related to causal-explanatory sentences like "The power went out because the lightning struck." A popular broadly Davidsonian view would be that the explanatory sentence is always an abstraction from the particular matters, and hence does not inquire, to stick with the example, about a particular striking and a particular power outage, although it does entail that there is some such event.[183] Another view holds that explanatory sentences and causal sentences can come apart entirely (such as when we explain the fact that the match did not light in virtue of the fact that it was not struck – seemingly no corresponding causal sentence there!).[184]

Here I shall mostly bracket these views because my question is not generally about the relation between causation and causal explanation. It is instead about the relation between unified explanation (which I take to be causal; but dissenting views are discussed below) and the causal matters which underlie it. More specifically, since what makes explanations unified is that they explain relational facts, an obvious question is how these relational facts relate back to the particular causal matters of the case. In the following, I want to offer two pictures of how the relation works, based on two pictures of the causal matters we find in the universe.

The Humean picture starts with a view of the causal stuff of the universe according to which no relational stuff exists – only separate particular events, distributed over space-time points, are available. It thus concludes that the relational fact cannot have a worldly analogue. Rather, we can understand it as picking out certain ways in which the particular worldly matters pattern, ways which are best expressed in modal statements.

The Aristotelian picture starts with the view that there is genuinely relational causal stuff – causal processes that is. It can thus represent the relational fact as corresponding to real relations in reality – causal processes that link up items of non-accidental relationships.

[183] The view shows up in several of Davidson's major works, including Davidson (1967a, 1967b, 1969, 1970a, 1970b).

[184] See especially Steward (1997), ch 5.

These two pictures about unified explanations, offer correspondingly different pictures on the metaphysics of non-accidentality, pictures that will prove helpful to fully understand the ambitions of this thesis and the issues I have discussed.

Because I have identified unified explanations with explanations that rely on the manifestation of dispositional properties, I will present the pictures just sketched as views about the metaphysics of the manifestation relation. It will be obvious how the relevant views belong to larger picture about the nature of non-accidentality.

Before I present the relevant pictures, let me provide a brief overview of approaches to the manifestation-relation. The manifestation-relation is rarely treated explicitly in philosophy, but we can loosely class stances, often implicit in treatments of deviance or dispositions, into four broad categories:

i. Primitivist or 'quietist' accounts, which claim either that 'manifestation' is conceptually primitive or that it 'up to the metaphysicians' to provide an account (see for example Sosa 2017, 136; Turri 2011, 7).
ii. Reductive accounts, often developed against the background of a Humean metaphysics, which roughly divide into accounts that either emphasize the disposition's role in laws or its role as a provider of modal patterns (Stoecker 2003, Smith 1997, 2009).
iii. Aristotelian accounts which take dispositions and manifestations to share an especially strong metaphysical link, often spelled out in terms of *causal processes* conceived as irreducible aspects of the world. (Mayr 2011, Stout 1996, ch.3, Stout 2005, ch.6, Stout 2010).
iv. Non-causal accounts (Mantel 2018, in some moods).

I will mainly concentrate on ii and iii in the following. Before I do, let me say a few things about i and iv.

8. Primitivism and Non-Causalism

As I have already pointed out above, primitivist solutions are unsatisfactory because they fail both the Job Description Requirement and the Non-Accidentality Requirement. They neither assign a clear explanatory role to the dispositional property in explaining the outcome, nor do they explain why the involvement of the dispositional property disperses the impression of accidentality. There is a tendency, amongst those who seem to adhere to primitivist views, to outsource the work of spelling out exactly what the

manifestation relation consists in, to philosophers of metaphysics. But this move creates an artificial distinction between the metaphysics of manifestations and their other uses. Using the notion of an exercise/manifestation to solve the problem of deviance just *is* a metaphysical project. And as the problem of deviance for the manifestation-relation shows, it is a project that can't be successfully executed without saying more about the manifestation-relation than the primitivist is willing to do.

What about the suggestion that the manifestation-relation can be spelled out non-causally? In her book *Determined by Reasons* Susanne Mantel presents what she calls "a competence account of acting for normative reasons". The competence in question is to track normative features in the world and produce matching behaviour. Mantel's normative competence is therefore almost identical to what I called the capacity to respond to reasons. Moreover, Mantel also agrees that deviant causal chains can be eliminated by appealing to the notion of a manifestation. However, somewhat surprisingly, later in the book she alleges that dispositional explanations are non-causal in nature. The argument is not entirely clear, but I take it that her line of reasoning goes like this:

Dispositional explanations of causal phenomena are indispensable because of the possibility of deviant causal chains. However, there are clearly instances of non-causal dispositional explanations. So the additional explanatory work done in deviance cases – and thereby in reasons-explanations of actions – must also be non-causal.

Mantel's examples for non-causal dispositional explanations are a tsunami causing the sirens to wail and a squirrel collecting nuts for the winter. These explanations are non-causal because their triggering events lie in the future:

> […] certainly the fact that there will be a tsunami, or the future event of there being a tsunami, does not cause the sirens to wail. Nor does the winter cause the squirrel to collect nuts. (Mantel 2018,168)

Mantel chooses these types of example for a reason. She thinks that the type of normative tracking competence she is after sometimes reacts to indicators of the features tracked rather than those features themselves. There appears to be some plausibility in claiming the same thing about the squirrel and the siren examples. The future event of the tsunami explains the wailing of the sirens because the sirens react to present indicators of the tsunami and the future event of the beginning of winter explains the behaviour of the squirrel because the squirrel is reacting to indicators of winter. Mantel tries to spell out the job description of the disposition as follows:

> Note that, although they are non-causal, these explanations have informative causal implications, as they exclude certain causal histories. If the sirens are

> wailing because there will be a tsunami and they manifest their disposition to wail before there is one, their wailing is not caused by some malfunctioning. Dispositional explanations are nevertheless distinct from causal explanations insofar as they are less informative about the particular cause than causal explanations are, but they are more informative insofar as they imply that the causal process must not have involved a malfunctioning of the disposition. (Mantel 2018, 146)

Technically, Mantel's view is compatible with the Doubly Explanatory Account. This is because the Doubly Explanatory Account does not specify that the dispositional property must *causally* explain why C caused o to φ. The view can therefore be conjoined with the sentiment expressed in the quote, which is that the fact that o had the disposition to φ in C provides non-causal additional information about the causal process 'C --> φ-ing'.

However, we have to be careful about what type of information is acceptable if we want to stick to the actual sequence spirit this thesis has been pursuing. For it would be easy to assume that if the additional information isn't about the particular cause and the particular φ-ing in question, then it is information about how the causal process fits into a larger modal pattern of very similar causes and φ-ings in the relevant class of possible worlds. If this were to be our preferred reading of Mantel (although she herself seems to reject modalist models, see Mantel 2018, 58), then we would have conceded defeat to modalism about reasons-responsiveness. Under this new modalist reading, even though we can understand exercising the RR capacity in terms of a special explanation, this explanation itself can only be established by adding modal information to a causal connection. Thus, the thought is that the actual world features that are pertinent for an agent's actually responding to reasons are not enough to get to a unified explanation of action. The actual world is impoverished in this respect, as it were. To bring out this worry more clearly, let me now look at the Humean and Aristotelian approaches, the former being (more or less) explicitly committed to the kind of 'poverty of actuality' thought just sketched.

9. Humeanism

Humeanism, in its most general form[185], is the thesis that there are no necessary connections between distinct existences. Specifically, for the things that *happen* in the

[185] Humeanism is often discussed as Humeanism about the laws of nature – the idea that the laws of nature are mere generalisations over the basic separate stuff that forms the mosaic. See Bhogal (forthcoming) for discussion.

universe, the thesis holds that all there is are distinct separate particular events. No causal necessitation or other relations obtain between these events, ultimately. These relations must instead ultimately be shown to disappear in pixilation once we zoom into the picture close enough. This pointillist picture of the universe is perhaps most famously given expression by David Lewis, who summarises it as follows:

> Humean supervenience...is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another....We have geometry: a system of external relations of spatiotemporal distance between points. Maybe points of space-time itself, maybe point-sized bits of matter or aether or fields, maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that. (Lewis1986b, pp.ix–x)

This reductive attitude will afflict all notions that express relations between particular matters, including of course orthonomy notions (a broadly Lewisian treatment of reasons-responsiveness was my target in chapter 2). Hence, when we provide the relational information coincidence questions are asking for, within the Humean picture this information must ultimately be about distinct causal matters. For reasons-responsiveness, what we are asking about is the connection between the fact that the agent φ-ed and the fact that the φ-ing matched with a reason. And we are answering this question by pointing out the agent exercised their capacity to translate this reason into φ-ings. We are thereby conceiving of the reason as at once the cause and the dispositional trigger for the φ-ing. In the Humean picture, this answer cannot be ultimately about some type of real dispositional influence – a dispositional connection between reason and action, or a ratio-causing, as I called it in the last chapter. For in the end, all the metaphysics will provide us with are two distinct existences: the reason and the action. The Humean will thus have to reduce the role of the dispositional property to facts about these distinct existences.

The best example for how the Humean can achieve this is Stoecker (1993, 2001, 2003) who thinks actions[1] are "actualisations of causal powers".[186] Dispositions in turn play the role of "boundary conditions which allow for the occurring causes to take affect" (Stoecker 2003, 297). Stoecker summarises the approach as follows:

[186] For Stoecker, causal powers are distinct from dispositions in that the tendencies of dispositions only concern what happens to the object bearing the disposition. Stoecker's use of "causal power" is thus roughly the same as my "disposition" or "capacity" whereas I would classify his "disposition" as a special class of causal power. Stoecker may sometimes sound like he rejects Humeanism (which sits uncomfortably with causal power talk), but much of what he says indicates a heavily reductive stance on causal powers.

> […] dispositions like being rotten are partially defined by their explanatory role: something is rotten only if it is in a state that could explain why occurrences that usually do not shatter things of the same kind (that don't have the disposition) can cause its collapse. To say that a bridge collapsed because it was rotten is to say that one could expect it to be destroyed by comparatively feeble impulses (like a truck crossing it), and that one such factor actually occurred and caused its collapse. This is what we mean when we say that the collapse was the actualization of the bridge's disposition of being rotten. (Stoecker 2003, 298)

So far, this looks compatible with the Doubly Explanatory Account. The crucial question here is how Stoecker aims to spell out the explanatory role of the dispositional property. The idea of dispositions as background conditions is telling here. It seems to me that something cannot 'allow' another thing to take affect if it does not exert some kind of causal influence or at the very least plays a causal role. Allowing refers to something that takes place on the causal level, not the causal-explanatory level alone, in Strawsonian terms. But on a Humean picture, the dispositional property can't exert its own dispositional influence. This is why the typical (already familiar) Humean response is to spell out the role of dispositional properties in explanations with appeal to Jackson and Pettit's idea of 'program explanations'.

Program explanations are premised on the thought that a feature may be causally *relevant* – and thus important in a full causal explanation of some phenomenon – without being causally *productive* of that phenomenon. We seem to encounter explanations that contain reference to features that are relevant in this way a lot. To take one of Jackson and Pettit's examples (originally from Putnam): We explain the fact that a square peg does not fit into a round hole equal in diameter to the side of the peg by referring to the property of squareness. And this seems to be an informative explanation of the fact that the peg does not fit. But the actual failure to fit is produced by the microphysical and spatiotemporal properties of the peg (and the hole).

According to Jackson and Pettit, a property can be causally relevant without being causally productive in this way by, as they call it, "programming" for the actually productive property. Most of the time, programming is spelled out by Jackson and Pettit as a type of realisation of properties. The squareness is causally explanatory of the failure to fit because its realisation ensures that the relevant causally productive properties are realized. Or, more generally:

> The realization of [M] ensures (…), that a crucial productive property is realized and (…), that the [effect] event.., occurs. [M] does not figure in the productive process leading to the event but it more or less ensures that a property-instance which is required for that process does figure [I]ts realization programs for the

> appearance of the productive property and.., for the event produced (Jackson & Pettit 1990a, 114).

Program explanations provide us with exclusive information by situating their corresponding process-information in a larger pattern. They do this by showing how the actual causal history of the phenomenon explained is related to sets of possible causal histories. I.e., program explanations provide *modal* information of a certain type (Jackson and Pettit ibid).

The peg example provides a useful way to illustrate this. The actual causes of the failure to fit are certain microphysical properties of the peg. But presumably, a whole range of similarly shaped objects would also not have fit in similarly shaped holes. All of these situations have something in common, namely that in each the actual causes of the failure to fit are realized because the squareness is realized. On this account, when we say that the peg does not fit because it is square, this is informative precisely because it is the squareness that remains invariant under many possible variations. It is the structure that emerges from a whole host of cases. The explanation that mentions squareness is also the more interesting one in this case. The many different ways in which the particular microproperties might cause a particular failure to fit seem negligible and too specific in this case. What gives us larger insight into the problem are the general geometrical shapes in play – and those shapes remain the same under modal variation. As J and P put it:

> We can express the basic idea behind a programme explanation in terms of what remains constant under variation. Suppose state a caused state b. Variations on a, say, a', a",(…), would have caused variations on b, say b', b", (…)respectively. It may be that if the a i share a property P, the b i would share a property Q: keep P constant among the actual and possible causes, and Q remains constant among the actual and possible effects(…). [I]n such a case P causally explains Q by programming it, even though it may be that P does not produce Q (Jackson & Pettit 1988, 394).

This picture now feeds back into Stoecker's account of the manifestation relation. For as I said, the program explanation picture allows for the dispositional property to show up on the causal level without any metaphysically problematic consequences. No specific 'dispositional influence' on the causal level has to be assumed. What it means for the breaking to be a manifestation of a glass' disposition to break if struck is for the striking to cause the breaking and for this (event) causal process falling within a pattern of similar causes causing similar outcomes across a modal subspace. Just like in the peg example the dispositional property programs for certain microstructures and so is causally

relevant, albeit not causally productive (the stone and base of the glass are), of the breaking of the glass.

The same of course goes for responding to reasons. On the Humean picture, to say that some φ-ing counts as the exercise of the capacity to respond to reasons for φ-ing is for the reason to cause that φ-ing and for the causal process to fall in a modal pattern of similar reasons causing similar φ-ings across a modal subspace. We already know this formulation of reasons-responsiveness from chapter 2 and 3. The Humean picture here converges with modalist accounts of responding to reasons.

We can also now see how the account pays service to the Humean picture of a mosaic of distinct existences. It follows the idea that we can reduce the manifestation relation to a pattern of distinct events. In order to make available the typical exercise-explanation corresponding to the exercise or manifestation of a dispositional property, the account admits, we need to appeal to extra information. But this information will (have to) be about the same resources the Humean finds acceptable: separate particular events. The additional information expressed in mentioning the dispositional property is then about how this metaphysically acceptable stuff patterns not just actually but modally. This is the expression of the thought that actuality is impoverished, that it can't muster the resources to explain the robustness of non-accidental connections.

It should be obvious that the spirit of my thesis – the ambition to understand non-accidentality not in modal but in explanatory terms – clashes with this way of spelling out the explanatory contribution of dispositional properties. In a nutshell, my argument in this thesis was precisely that modal patterns are neither a good guide nor a good metaphysical bedrock for non-accidental relationships. My central argument – which explains the recalcitrant counterexamples discussed in chapter 2 – was that modalism must ultimately treat non-accidental relationships compositely. The Humean picture provides a metaphysical model for this composite treatment. If all there is in the world are just separate particular facts, then all facts must ultimately either be non-relational or be composites of non-relational facts. Modalism is congenial with the Humean picture precisely because it honours the distinctness of existences while promising an innocent analysis of pesky relational matters – such as causation and orthonomous relatonships. But it is precisely here where the modalist picture falters. For these notions, it seems my arguments show, require a stronger relational metaphysics, one that allows there to be fundamental relational facts.

There is an illuminating similarity here between this discussion and the old debate on occasionalism.[187] The similarity is this: what seems strange about occasionalism is that it

[187] See Nadler (2010) for an overview.

proposes that events are wholly distinct existences, connected solely by the fact that god's will imposes on them a certain sequential order, sometimes mistaken by panicked philosophers as causal, or worse, necessary connections. This idea seems strange most fundamentally, I think, because the supposition that causes and effects are connected in a mere conjunction mediated by god makes the connection between cause and effect seem too accidental. What I have suggested seems strange about modalist accounts of non-accidentality is something very similar: If all that connects causes and effects, reasons and responses, is a certain sort of modal pattern, then we have not improved occasionalism substantially. We have merely added the domain of possibilia for the abstraction to range over.

The central spirit of this thesis then points to an underlying Anti-Humean metaphysics. In order to explore such a metaphysics, I will in the next section look at Rowland Stout's Aristotelian account.

10. Aristotelianism

There is an entirely different outlook on the exercise of capacities available, which sharply contrasts with the Humean modal pattern view. The kernel of these Aristotelian views[188] (I'm following Stout 2010 with this label) is that exercises of dispositions are sui generis processes, not reducible to chains of events. For Stout (1996, 2005, 2010), the relationship between disposition and exercise is even more intimate. The literal usage of the notion of 'manifestation' implies that manifestations are versions of the property/object manifested. Dispositions literally are their manifestations in this picture because Stout thinks of them as "dynamic states" of dispositional systems. The view is complicated (and obscure in places). An attempt at brief reconstruction:

Dispositional systems come with sets of very specific operational conditions on Stouts view. If the operational conditions for the relevant potentiality are satisfied, then the whole manifestation-process is present at every temporal stage. The process of decay of an apple, roughly, means that that there is some potentiality of decay that comes with characteristic operational conditions and these characteristic conditions are satisfied such that at every temporal stage of the apple's decay the apple is decaying – so what is present in the particular temporal region is not only a temporal part of a process, but the whole process. The idea is that this sets 'Aristotelian' processes apart from what Stout calls Russellian processes, which are chains of events. These processes count as

[188] I am focussing on Stout here, who focusses on processes, but the label is meant to encompass all views which posit a kind of dispositional causation, such as Heil and Martin (1998), Heil and Martin (1999), Mumford and Anjum (2011), Marcus (2012), and Williams (2019).

particulars, parts of the causal inventory of the world (as opposed to the kind of patterned generalisations the Humean picture identifies).[189]

Stout summarises the idea as follows:

> To clarify the contrast between the two conceptions of a process, consider this process of salt dissolving in water. On the Russellian view this process consists of a causally connected continuous chain of stages from solid crystals of salt being immersed in the water to all the salt being dissolved in the water. On the Aristotelian view the process is identified with the realization of the potentiality of the salt to dissolve in water. In other words it is the realization of the various operational conditions of the water–salt dissolving system. The state of fulfilment of the potentiality of salt to dissolve is precisely the process of the salt dissolving in water. On this view a process is a state – a dynamic state – not a series of events. (Stout 2005, 89)

Another way to grasp the intuition behind this view is to say that when a potentiality is actualized, this means a causal system is active. It is undergoing an activity. Of this intuitive notion of an activity, too, we would say that it is present in its entirety at every temporal stage of the activity. Equally we would prima facie say of it that it cannot be reduced to a series of event-causal steps. Some behaviours of objects lend themselves better to eliciting this idea. One kind of behaviour that I think illustrates Stout's concept quite well is the glowing of a ball. When the ball glows, there is a potentiality of the ball with very characteristic operational conditions. When these conditions are satisfied, the ball is glowing. It is manifesting its potentiality to glow. But the example is prima facie quite resistant to a reductive analysis according to which the causal activity of the ball consists in a series of glowing events. This resistance is not only due to the continuous nature of the glowing activity, but also to the natural idea that each point in the time during which the glowing takes place, the ball's capacity to glow is, at it were, fully realised. And by fully realised we mean that the entire glowing phenomenon is present at each stage of the glowing.[190] This latter intuition is of course exactly Stout's own, namely that the whole particular process is present at each point in time at which the operational conditions are satisfied.

The outcome of this metaphysical picture ties into the problems with coincidence for the Humean picture. First, each dispositional system comes with its own dynamic state, a process irreducible to a concatenation of events. Insofar as we consider reasons-

[189] Stout (1996), 155–163 holds that these processes are non-causal.

[190] Stout holds that processes are continuants.

responsiveness as such a dispositional system (or more plausibly several interrelated systems), this means it becomes very hard to imagine cases of accidentality in which a reason which rationalises the agents action also causes it, because in the Aristotelian picture there aren't two facts here in the first place. There is no causing and rationalising. There is just a rational-causing (Marcus 2012). There is no having reasons and a separate acting, there is literally just an acting-for-reasons process. The Aristotelian view closes the event gaps which allow the possibility of accidentality on a metaphysical level in the first place[191], but at the cost of introducing real connections into the fabric of the world – precisely the kinds of local necessities that the Humean picture is trying to avoid. [192]

These considerations show that the Aristotelian metaphysics sketched here is much more appropriate to the arguments I have advanced in this thesis. My central argument against modalism about non-accidentality is that non-accidental relationships cannot be decomposed into separate parts. Stouts picture offers a way to represent this insight on the metaphysical level: they cannot be decomposed, it suggests, because what metaphysically grounds non-accidental relations are sui generis processes irreducible to separately existing events. Let me put this in the explanatory language I used in the last chapter. The Aristotelian view is congenial to my view on exercising capacities and non-accidentality because I illuminated these notions in terms of what I called unified explanations. The central mark of a unified explanation of [p&q], I pointed out, is that we are seeking relational information about [p&q]. That is, we are seeking to explain the relational fact that [p&q], not just the fact that p and the fact that q (this is why the 'because' operator does not work distributively in these explanations). But if we ultimately metaphysically link p and q to underlying distinct existences – events, most paradigmatically – then we will never be able to hone in on the proper relational fact (at least not on the metaphysical level, more on this in the next chapter). If we instead correlate [p&q] with an underlying sui generis 'p-q' process, then we immediately get

[191] Stout says: "You cannot interpolate a deviant causal chain into the stages of a teleological process and still have the same teleological process happening. This is because the process with a deviant causal chain will have different characteristic stages. Even if it results in the same thing on this occasion, the presence of the deviant chain will mean that there will be some situations in which the augmented process is happening where it does not result in what it should according to the means-end justification. This is by contrast with teleological causal chains. You can always interpolate a deviant causal chain into a teleological causal chain and still have the same teleological causal chain. This is why the problem of deviant causal chains can only be solved with an account of processes as distinct from causal chains." (Stout 1996, 91)

[192] Another issue that will be seen as a burden of the account is that it would appear to commit us to 'meta-causation' i.e. the notion that facts about causal relations can themselves be caused. This is because on the account, the disposition explaining why C caused o to φ corresponds to the dispositions being causally involved in the production of the 'C → φ'- process. I think we should use a notion of causal sustainment to spell this contribution out. But some (Dasgupta 2014, 568, fn 23; deRoset 2013, 19; Schaffer 2017, 19-20) would still find the general idea of meta-causation mysterious. Notably though, a variety of different models in philosophy are committed to it, for example Hitchcock (1996); Koons (1998); Barden (2014) – or even outright embrace it, like Needham (1988), 215 – 216; Mellor (1995), 106.

the relevant relational fact. It is a fact about the process p-q, not a composite of the fact p (about event 1) and the fact that q (about event 2).

This was of course only a brief overview of the kinds of metaphysical pictures the Doubly Explanatory Account might be thought to be compatible with. I lack the space in this thesis to develop the metaphysics further. But this should be fine, as part of the point of developing an explanationist account of x is that we can learn a lot of philosophically useful information about x just in terms of x's explanatory properties – irrespective of what metaphysics ultimately underpins these properties.

With what I have done in this chapter, my account of actually responding to reasons is complete (or as complete as possible, at least). Let us then take stock.

General Conclusion

In this thesis, I have tried to lay the groundwork for a novel way to understand the orthonomous relationship that connects reason and action when the agent responds to reasons – the relationship that in turn grounds free agency. I have pointed out that all attempts to understand the relationship modally are beset by what seems to me insurmountable problems, problems that can moreover clearly be traced to the fact that we have explanatory intuitions about the orthonomous relationship.

The clash between the default modal analysis and our intuitions about explanatory connections concerns, I have proposed, the type of relationship orthonomous relationships are at their core – non-accidental relationships, that is. My strategy accordingly has been to develop an account of responding to reasons that captures the non-accidentality of the notion of a response in explanatory terms.

I executed this strategy in two steps. First, I argued that responding to reasons involves the *exercise* of the capacity to respond to reasons, and I proposed that we can gain an initial understanding of exercising capacities by attending to the fact that they are involved in exercise-explanations. But I left the notion of an exercise unanalysed beyond that. I showed that, taking the notion of an exercise as basic, we can develop a fruitful account of responding to reasons that allows us to see how the phenomenon is unified across cases of error and success.

The second step was to explore that explanatory mechanics of exercise-explanations. At the core of this second step lies my fully non-modal, explanationist understanding of non-accidentality in terms of unified explanations – special explanations of relational facts that cannot be decomposed into separate independent non-relational facts. Not only does this account explain our general intuitions about coincidences (coincidences, intuitively, are unrelatedly co-instantiated facts), it gains immense additional support from the fact that it offers a template foe unifying a whole range of problem cases that have not so far been recognized as interestingly connected. Masking, finking, and mimicking cases have been discussed mainly as obstacles for a conditional analysis of dispositional properties. Cases of epistemic coincidence have mainly occupied epistemology alone. Discussion of cases of causal deviance has been confined, for the most part, to the philosophy of action (although especially writers in the nineties had already recognized ho far-reaching their impact really is). My explanationist approach –

and the general perspective taken in this thesis – show that these three classes of cases are intimately related. They are all, at base, instances of problems with accidentality. More precisely, they are all cases in which we fail to explain certain relational facts in a unified way.

Finally, I reintegrated my general explanationist account of non-accidentality with the exercise-based account developed in the first step. Exercise-explanations indispensably involve reference to the relevant dispositional property. To give an explanation of a φ-ing in terms of this property is to explain, by reference to the disposition, why the causes of the φ-ing led to that φ-ing. It is to explain, in other words, the causal role of something in terms of its role as the trigger for the relevant dispositional property. This is how exercise-explanations provide precisely the kind of unification required to disperse non-accidentality.

These components provide us with a new way of understanding responding to reasons. An action is a response to a reason, according to this view, if we can give a unified explanation of the action in terms of that reason – an explanation of the relational fact that the agent did what was the rational thing to do. Thus, we can understand responding to reasons and the essential non-accidentality it involves, as an explanatory phenomenon. And in turn, we can answer the question about the orthonomous relationship in explanatory terms, too. Actions and reasons are orthonomously related when the fact that the reason caused the action can be explained in terms of the exercise of the agent's rational powers. No reference to alternative possibilities is required. This is an actual actual-sequence approach.

I believe I have already shown throughout this thesis that my project has considerable potential for illuminating systematisation of philosophical problems hitherto thought unrelated and for offering new ways of conceiving of such central notions as knowledge, moral worth, and rule-following. I believe it also has more far-reaching consequences. It forces us to adopt a new philosophy of mind and action, for example, one in which the notion of the exercise of a capacity takes centre stage. However, here I want to point out two lacunae in my defence of the explanationist project that I would like to understand as additional avenues for future research. Both concern the way in which alternative possibilities might be thought to make a comeback in some parts of the project. All I will do here is briefly explain the spirit of the problem and hint at a solution.

First, I have omitted (except hints in some footnotes) discussion of the type of case which is now known as 'Gettier-case' (I did discuss the original Russel examples, which do not cause problems do my view). In the typical version of such cases, a person is travelling through a landscape full of fake barns. Looking, by accident, at the only real barn in the

vicinity, the person forms the belief that the object in front of them is a barn. Epistemologists in the grip of safety conditions on knowledge have thought that this sort of luckiness undermines any claims to knowledge about the object in front of them the person might otherwise have. And these cases do seem relevant to the explanationist approach. For it might seem like the explanatory connection between the person's perceptual reasons and their belief is not impinged, and yet the perceived luckiness of the belief seems to undermine their claim to knowledge.

Allow me to insert an honest comment about my own position first. I have never fully felt the pull of the anti-knowledge intuitions about these cases. It seems to be that rather than motivating a safety condition, anti-knowledge intuitions about the barn cases are *motivated by* an antecedent adherence to such conditions. In short, I think in these cases the subject *does* know that the object in front of them is a barn, and they *are* responsive to reasons for their belief. One possibility for future research is a novel attempt at vindicating such intuitions in terms of facts about the flawless exercise of the agent's capacities in these cases. However, it is obviously free to explanationists to explore different strategies. Improved clarity about the concept of non-accidentality will allow us to map its relation to other, distinct forms of luckiness, allowing us to see in just what ways they interfere and/or intersect with explanations of the relevant facts.

Second, I have not included in this thesis an explicit discussion of the idea that modal properties – properties than can only be spelled out in terms of alternative possibilities - can themselves be explanatory. If modal properties can be involved in explanation, this might spell trouble for the distinction between modalism and explanationism. The best way to demonstrate how the trouble might be thought to arise is to look at a twin case that Sartorio (2016) discusses extensively in her actual-sequence approach to free agency:

> **Not All Roads Lead to Violent Act**: Jones has some innate violent tendencies. Smith has just done something that Jones found very upsetting. Absent any interventions or distractions, Jones knows that his current train of thought would result in the triggering of some irresistible urge to harm Smith. After some helpful sessions with his therapist, Jones knows that he can create a successful distraction. His love for logical puzzles is so powerful that, if he starts working on one of those puzzles, he'll be so immersed in it that the desire to harm Smith will pass. Still, Jones decides not to engage in the distraction. As a result, the violent urge is triggered and Jones harms Smith.
>
> **All Roads Lead to Violent Act**: Everything is the same as in Not All Roads Lead to Violent Act (in particular, Jones knows that his current train of thought will

> trigger the irresistible urge to harm Smith unless he creates a successful distraction) except that, in this case, unbeknownst to Jones, there is no way for him to create a successful distraction. Some evil neuroscientist has tinkered with his brain in such a way that working on the logical puzzle would not have had the expected effect. (Sartorio 2016, 62 -63)

The worry expressed in comparing these two cases is that they are identical in terms of their actual sequences but not in their respective action's freedom-status. The agent in Not All Roads is clearly in control of their action (beating up Smith) whereas it seems at least questionable whether the agent in All Roads has the same status. At the very least their control over their action seems diminished. But at first sight, there is nothing in the actual sequence that could account for that difference. A salient modal difference however is that the agent in Not All Roads appears to keep the ability to do otherwise.

This constellation of judgments prima facie suggests that free agency cannot be grounded solely in what explains the actual action but must be grounded at least in part by what would have explained the non-actual action. Considerations like this in fact underlie some of the issues I discussed in chapter 4, which dealt with cases of error. I argued there that there is an indispensable level of evaluation that applies to agent's only if they are already exercising the relevant capacities. But what about cases in which agents do not exercise their capacity to φ, but nonetheless *could have* φ-ed? Surely their failures are in part explained by the fact that they did not do what they could have done. Or more generally: When you do what you ought to have done because you recognized that you ought to do it, then your doing is explained by the fact that you ought to have done it. But if you fail to do what you ought to have done, then your wrong-doing is in part explained by the fact that you could have done something else.

These considerations then point to a larger problem. I have focussed in this thesis on explaining the action that is the rational thing to do. And I have said about irrational actions only that they will sometimes count as exercises of the capacity to respond to reasons, namely when they are manifestations of the relevant capacity not functioning well. But the background puzzle is how we can explain, via the capacity to respond to reasons, actions that are irrational and so *aren't* responses to reasons, in some sense. If we want to stick with an actual-sequence approach, we better not appeal to the fact that the agent could have done the rational thing.

I believe the way forward is to explore the way in which we can often explain divergent instances of something by appeal to what they are divergent *from*. We explain them under the description of being divergences, that is. We can explain exceptions to rules, for example, in part by appealing to the rule itself. I propose that we should think about

explaining irrational actions – those irrational actions that are still exercises of the capacity to respond to reasons, that is – in a similar way. Tellingly, amongst those instances to which we ordinarily apply divergence explanations are malfunctions of systems. I have already characterised irrational actions as malfunctions of the capacity to respond to reasons in this thesis. So to explore the explanatory profile of malfunctions is the natural next step for the project. At the face of it, explanations in terms of malfunctions do not require us to cite the ability to do otherwise. Hence, the possibility of malfunction explanations opens up a way for spotting a purely explanatory difference between All Roads and Not All Roads above. The surface difference between the two is that Jones exercises his capacity to respond to reasons in Not All Roads, but not in All Roads. The deeper difference corresponding to that is that we have available an explanation in terms of a malfunction for Jones in Not All Roads that isn't available for Jones in All Roads.

This branch of the explanationist project, when fully developed, then has the potential to unlock new ways of thinking about the typical phenomena of practical irrationality – chief among them weakness of the will.

References

Aguilar, Jesús H. (2012). Basic causal deviance, action repertoires, and reliability. *Philosophical Issues* 22 (1),1-19.

Altham, J. E. J. (1986). The Legacy of Emotivism. In Graham Frank Macdonald & Crispin Wright (eds.), *Fact, science and morality: essays on A.J. Ayer's Language, Truth and Logic*. Oxford: Basil Blackwell, 275-288.

Alvarez, M. & Hyman, J. (1998). Agents and their actions. *Philosophy,* 73 (2), 219-245.

Alvarez, M. (2008). Reasons and the ambiguity of "Belief". *Philosophical Explorations*,11, 53–65.

Alvarez, M. (2009a). Actions, thought-Experiments and the 'principle of alternate possibilities'. Australasian *Journal of Philosophy*, 87, 61-81.

Alvarez, M. (2009b). Acting Intentionally and Acting for a Reason. *Inquiry: An Interdisciplinary Journal of Philosophy,* 52 (3), 293-305.

Alvarez, M. (2010). *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press.

Alvarez, M. (2013) Agency and two-way powers. *Proc Aristotelian Soc* 113,101–121.

Alvarez, M. (2017). Are Character Traits Dispositions? *Royal Institute of Philosophy Supplement* 80, 69-86.

Alvarez, M. (2018). Reasons for action, acting for reasons, and rationality. *Synthese* 195 (8), 3293-3310.

Anscombe, G.E.M. (1963). *Intention* (2nd Edition). Harvard University Press.

Arpaly, N. (2000). On Acting Rationally Against One's Best Judgment. *Ethics* 110 (3), 488-513.

Arpaly, N. (2002). Moral Worth. *The Journal of Philosophy*, 99(5), 223–245.

Arpaly, N. (2003). *Unprincipled Virtue*. New York, NY: Oxford University Press.

Arpaly, N. (2006). *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton University Press.

Arpaly, N. & Schroeder, T. (2014). *In Praise of Desire*. Oxford University Press.

Audi, R. (1990). Weakness of will and rational action. *Australasian Journal of Philosophy* 68 (3), 270 – 281.

Audi, R. (1993). *The Structure of Justification*. Cambridge University Press.

Austin, J.L., (1956). Ifs and Cans. *Proceedings of The British Academy* 42, 107-132.

Ayer, A. J., 1954. Freedom and Necessity. In his *Philosophical Essays*, New York: St. Martin's Press, 3–20; reprinted in Watson (ed.), 1982,15–23.

Bach, K. (1980). Actions are not events. *Mind* 89 (353), 114-120.

Barnden, J. (2014). Running into Consciousness. *Journal of Consciousness Studies,* 21 (5-6), 33-56.

Barnett, D. (2009). The myth of the categorical counterfactual. *Philosophical Studies,* 144 (2), 281 - 296.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 421–441.

Berofsky, B. (2002). Ifs, cans, and free will: The issues. In Kane, R.H. (ed.), *The Oxford Handbook of Free Will*. Oxford: Oxford University Press, 181-201.

Bhogal, Harjit (2020). Coincidences and the Grain of Explanation. *Philosophy and Phenomenological Research* 100 (3), 677-694

Bhogal, Harjit (forthcoming). Humeanism about Laws of Nature. *Philosophy Compass*.

Bird, A. (1998). Dispositions and antidotes. *Philosophical Quarterly,* 48, 227–234.

Bishop, J. (1989). *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press.

Black, T., and P. Murphy. (2007). In Defense of Sensitivity. *Synthese,* 154 (1), 53–71.

Black, T. (2008). Defending a Sensitive Neo-Moorean Invariantism. In: V. F. Hendricks and D. Pritchard (eds), *New Waves in Epistemology*. Palgrave Macmillan, 8–27.

Block, N. (1990). Can the Mind Change the World? In George Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press, Cambridge, 137-170.

Bogardus, T. & Perrin, W. (forthcoming). Knowledge is Believing Something Because It's True. *Episteme*,1-19.

Bonevac, D., Dever, J., & Sosa, D. (2006). The conditional fallacy. *Philosophical Review,* 115, 273–316.

Bradford, G. (2015). *Achievement*. Oxford University Press.

Brand, M. (1984). *Intending and Acting: Toward a Naturalized Action Theory*. Cambridge, MA: MIT Press.

Brink, D. O. & Nelkin, D. K. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility* 1, 284-313.

Broome, J. (2007). Does Rationality Consist in Responding Correctly to Reasons? *Journal of Moral Philosophy* 4 (3), 349-374.

Broome, J. (2008). Is rationality normative? *Disputatio* 2 (23), 161-178.

Broome, J. (2009). Motivation." *Theoria* 75 (2), 79–99.

Broome, J. (2013). *Rationality Through Reasoning*. Wiley-Blackwell.

Brueckner, A. (2002). Williamson on the primeness of knowing. *Analysis* 62 (3), 197–202.

Carr, D. (1979). The Logic of Knowing How and Ability. *Mind* 88, 394–409.

Carter, J. A. (2019). Exercising Abilities. *Synthese*, 1-15.

Chang, R. (2001a). Two Conceptions of Reasons for Action. *Philosophy and Phenomenological Research* 62 (2), 447-453.

Chang, R. (2001b). *Making Comparisons Count*. Routledge.

Chang, R. (2013). Grounding practical normativity: going hybrid. *Philosophical Studies* 164 (1),163-187

Chang, R. (2014). Practical Reasons: The problem of gridlock. In Barry Dainton & Howard Robinson (eds.), *The Bloomsbury Companion to Analytic Philosophy*. Continuum Publishing Corporation, 474-499.

Chang, R. (2020). Do We Have Normative Powers? *Aristotelian Society Supplementary Volume* 94 (1), 275-300.

Chisholm, R. (I964). The Descriptive Element in the Concept of Action, *Journal of Philosophy*, lxi, 613-625.

Chisholm, R. (1964). Human Freedom and the Self. *The Lindley Lectures*, Copyright by the Department of Philosophy, University of Kansas. Reprinted in Watson (ed.) 1982.

Choi, S. (2005). Dispositions and Mimickers. *Philosophical Studies,* 122 (2),183-188.

Choi, S. (2006). The Simple Vs. Reformed Conditional Analysis of Dispositions. *Synthese* 148 (2), 369-379.

Choi, S. (2012). Intrinsic Finks and Dispositional/Categorical Distinction, *Nous* 46 (2), 289-325.

Choi, S. (2009). The conditional analysis of dispositions and the intrinsic dispositions thesis. *Philosophy and Phenomenological Research*, 78, 568–590.

Clarke, R. (2008). Dispositions, Abilities to Act, and Free Will: The New Dispositionalism. *Mind* 118, 323-351.

Clarke, R. (2010). Opposing powers. *Philosophical Studies* 149 (2),153-160.

Clarke-Doane, J. (2012). Morality and Mathematics: The Evolutionary Challenge. *Ethics* 122 (2), 313–340.

Clarke-Doane, J. (2014). Moral Epistemology: The Mathematics Analogy. *Noûs* 48 (2), 238–55.

Clarke Doane. J. (2015). Justification and Explanation in Mathematics and Morality. In R. Shafer-Landau (ed), *Oxford Studies in Metaethics,* Vol. 10, Oxford University Press, 80–103.

Clarke-Doane, J. (2016). What Is the Benacerraf Problem? In F. Pataut, *New Perspectives on the Philosophy of Paul Benacerraf: Truth, Objects, Infinity*. Springer, 17–43.

Clark P. (2001). Velleman's Autonomism. *Ethics*, 111(3), 580-593.

Coffman, E.J. (2007). Thinking about Luck. *Synthese,* 158, 385-398.

Cohen, D. & Handfield, T. (2007). Finking Frankfurt. *Philosophical Studies* 135 (3), 363-374.

Comesaña, J. (2005). Unsafe Knowledge. *Synthese,* 146 (3), 395-404.

Comesaña, J. & McGrath, M. (2014). Having False Reasons. In Clayton Littlejohn & John Turri (eds.), *Epistemic Norms*. Oxford University Press, 59-80.

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of Neuroscience*. Oxford University Press, Clarendon Press.

Cross, T. (2005). What is a disposition? *Synthese*, 144, 321–341.

Cullity, G. (2019). Weighing Reasons. In Daniel Star (ed.), *The Oxford Handbook of Reasons and Normativity*. Oxford: Oxford University Press

Cunningham, J. (2019a). Is believing for a normative reason a composite condition? *Synthese* 196 (9), 3889-3910.

Cunningham, J. (2019b). The Formulation of Disjunctivism About ϕ-ing for a Reason. *Philosophical Quarterly* 69 (275), 235-257.

Cuneo, T., (2007), *The Normative Web. An Argument for Moral Realism*, Oxford: Oxford University Press.

Dancy, J., (1993). *Moral Reasons*. Oxford: Basil Blackwell

Dancy, J. (2000). *Practical Reality*. Oxford University Press

Dancy, J. (2004a). *Ethics Without Principles*, Oxford: Oxford University Press.

Dancy, J. (2004b). Two Ways of Explaining Actions: Jonathan Dancy. *Royal Institute of Philosophy Supplement* 55, 25-42.

Dancy, J. (2008). On how to act : disjunctively. In A. Haddock & F. Macpherson (eds.), *Disjunctivism: Perception, Action, Knowledge*. Oxford University Press, 262-282.

Dancy, J. (2014). On knowing one's own reasons. In C. Littlejohn & J. Turri (Eds.), *Epistemic norms, new essays on action, belief and assertion*. Oxford: Oxford University Press, 81–96

Dardis, A. (1993). Sunburn: Independence Conditions on Causal Relevance. *Philosophy and Phenomenological Research*, 53 (3), 577-598.

Dasgupta, S. (2014). The Possibility of Physicalism. *Journal of Philosophy*, 9-10, 557-592.

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy* 60 (23), 685, reprinted in Davidson 1980, 3–20

Davidson, D. (1969). The Individuation of Events. In N. Rescher et al. (eds.), *Essays in Honor of Carl G. Hempel*. D. Reidel, Dordrecht, 216-234.

Davidson, D. (1967a). The Logical Form of Action Sentences. In N. Rescher (ed.) *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press), reprinted in Davidson (1980).

Davidson, D. (1967b). Causal Relations. *Journal of Philosophy,* 64, 691-703. Reprinted in Davidson (1980), 149- 162.

Davidson, D. (1970a). Mental Events. In L. Foster and J. W. Swanson (eds.), *Experience and Theory*. The University of Massachusetts Press, Amherst. Reprinted in Davidson, 1980, 207-225

Davidson, D. (1970b). Events as Particulars. *Nous* 4, reprinted in Davidson (1980).

Davidson, D. (1970c). How Is Weakness of the Will Possible? In Davidson 1980, 21–42

Davidson, D. (1973). Freedom to act. In T. Honderich (ed.), *Essays on Freedom of Action*. Routledge.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

DeRose, K. (1995). Solving the Skeptical Problem. *The Philosophical Review,* 104 (1), 1–52.

DeRosset, L. (2013). Grounding Explanations. *Philosophers' Imprint*, 13.

Dretske, F. (1971). Conclusive Reasons. *Australasian Journal of Philosophy,* 49 (1), 1–22.

Dretske F. and Enc. B. (1984). Causal Theories of Knowledge. In P. French, T. Uehling, and H. Wettstein (eds.), *Midwest Studies in Philosophy,* 9 (1), 517-528.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, Massachusetts: MIT Press.

Dretske, F. (1992). The metaphysics of freedom. *Canadian Journal of Philosophy* 22 (1), 1-13.

Dretske, F. (1993). Reasons as Structural Causes of Behavior. In Heil and Mele (1993).

Dretske, Fred (2009). What must actions be for reasons to explain them? In Sandis, C. (ed.), *New Essays on the Explanation of Action*. Palgrave-Macmillan, 13-21.

Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism*. Oxford University Press UK.

Enç, B. (2003). *How We Act: Causes, Reasons, and Intentions*. Oxford University Press.

Fantl, J. and McGrath, M. (2009). *Knowledge in an Uncertain World*. Oxford: Oxford University Press.

Fantl, J. (2015). What Is It to Be Happy That P? *Ergo: An Open Access Journal of Philosophy,* 2.

Fara, M. (2001). Dispositions and their ascriptions. Dissertation, Princeton University, Princeton.

Fara, M. (2005). Dispositions and habituals. *Noûs,* 39 (1), 43–82.

Fara, M. (2008). Masked Abilities and Compatibilism. *Mind,* 117 (468), 843–65.

Faraci, D. (2019). Groundwork for an Explanationist Account of Epistemic Coincidence. *Philosophers' Imprint,* 19.

Faye, J. (1999). Explanation explained. *Synthese,* 120 (1), 61-75.

Fileva, I. (2016). Two Senses of "Why": Traits and Reasons in the Explanation of Action. In *Questions of Character*. Oxford University Press,182-202.

Fischer, J.M. (1982). Responsibility and Control. *Journal of Philosophy*, 89, 24–40.

Fischer, J.M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell.

Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility.* New York: Cambridge University Press.

Fischer, J.M. (2006). *My Way: Essays on Moral Responsibility*. New York, NY: Oxford University Press.

Fischer, J. M. (2008). Freedom, Foreknowledge, and Frankfurt: A Reply to Vihvelin. *Canadian Journal of Philosophy,* 38 (3), 327-342.

Fischer, J.M. (2012). *Deep Control: Essays on Free Will and Value*. New York, NY: Oxford University Press.

Fisher, Justin C. (2013). Dispositions, conditionals and auspicious circumstances. *Philosophical Studies,* 164 (2), 443-464.

Frankfurt, H. (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy*, 66, 829–839.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 68, 5–20.

Franklin, C. E. (2015). Everyone thinks that an ability to do otherwise is necessary for free will and moral responsibility. *Philosophical Studies,* 172 (8), 2091-2107.

Franklin, C. E. (2011). Masks, Abilities and Opportunities: Why the new dispositionalism cannot succeed. *Modern Schoolman*, 88, 89-103.

Fricker, E. (2009). Is knowing a state of mind? The case against. In Duncan Pritchard & Patrick Greenough (eds.), *Williamson on Knowledge*. Oxford: Oxford University Press.

Garfinkel, A. (1981). *Forms of explanation*. Yale University Press New Haven.

Gettier, E. L. (1963). Is Justified True Belief Knowledge? *Analysis*, 23 (6), 121–3.

Ginet, C. (1996). In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing. *Philosophical Perspectives,* 10, 403-17.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Ginet, C. (2006). Working With Fischer and Ravizza's Account of Moral Responsibility. *Journal of Ethics*, 10, 229–53.

Goldman, A. I. (1970). A Theory of Human Action. Englewood Cliffs, NJ: Prentice-Hall.

Goldman, A. I. (1967). A Causal Theory of Knowing. *Journal of Philosophy,* 64, 355-72.

Greco, J. (2010). *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge University Press.

Gregory, A. (2013). The Guise of Reasons. *American Philosophical Quarterly,* 50 (1), 63-72.

Gregory, Alex (2017). Might Desires Be Beliefs About Normative Reasons? In J. Deonna & F. Lauria (eds.), *The Nature of Desire*. Oxford University Press, 201-217.

Grice, H. P. 1962. 'The Causal Theory of Perception', *Proceedings of the Aristotelian Society, Supplementary Volume;* also in R. J. Swarz (ed.), *Perceiving, Sensing and Knowing* (Berkeley: University of California Press, 1965).

Grice, H. P. (2001). *Aspects of Reason*. Clarendon Press.

Gundersen, L. (2002). In defence of the conditional account of dispositions. *Synthese*, 130, 389–411.

Haji, I. (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York, NY: Oxford University Press.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall & Laurie Paul (eds.), *Causation and Counterfactuals*. MIT Press, 225-276.

Heering, D. (forthcoming). Actual Sequences, Frankfurt-Cases, and Non-Accidentality. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Heering, D. (2020). Intentionen, Misserfolg und die Ausübung von Fähigkeiten: Bemerkungen zu Agents' Abilities von Romy Jaster. *Zeitschrift für Philosophische Forschung,* 74 (3), 454-459.

Heil, J. Mele, A. (eds.) (1993). *Mental Causation*. Oxford: Clarendon Press.

Heil, J. (2003). *From an Ontological Point of View*. New York: Oxford University Press.

Heil J. & Martin C. (1999). The Ontological Turn. *Midwest Studies in Philosophy*, 34-60,

Herman, Barbara (1981). On the Value of Acting from the Motive of Duty. *The Philosophical Review*, 90(3), 359–382.

Heuer, U. (2004). Reasons for actions and desires. *Philosophical Studies*, 121 (1), 43–63.

Hiller, A. & Neta, R. (2007). Safety and epistemic luck. *Synthese*,158 (3), 303 - 313.

Hitchcock, C. (1996). A probabilistic theory of second order causation. *Erkenntnis*, 44 (3), 369 - 377.

Hobart, R. E. (1934). Free Will as Involving Indeterminism and Inconceivable Without It. *Mind*, 43, 1–27.

Hornsby, J. (2004). Agency and Actions. *Royal Institute of Philosophy Supplement,* 55,1-23.

Hornsby J. (2007). Knowledge in action. In A. Lesit (Ed), *Action in Context*. De Gruyter.

Hornsby, J. (2008). A disjunctive conception of acting for reasons. In A. Haddock & F. Macpherson (Eds.), *Disjunctivism: Perception, action, knowledge*. Oxford: Oxford University Press, 244–261.

Horwich P. (1982). *Probability and Evidence*. Cambridge University Press,

Howard, N. R. (forthcoming). One Desire Too Many. *Philosophy and Phenomenological Research*.

Hume, D. (1978). *An Enquiry Concerning Human Understanding*, P.H. Nidditch (ed.), Oxford: Clarendon Press.

Hunt, D. P. (2000). Moral Responsibility and Unavoidable Action. *Philosophical Studies,* 97, 195–227.

Hunt, D. P. (2005). Moral Responsibility and Buffered Alternatives. *Midwest Studies in Philosophy,* 29, 126–45.

Hyman, J. (2015). *Action Knowledge & Will*. Oxford University Press UK.

Hyman, J. (1999). How knowledge works. *The Philosophical Quarterly*,49, 433–451.

Ichikawa, J. (2011). Quantifiers, Knowledge, and Counterfactuals. *Philosophy and Phenomenological Research,* 82 (2), 287–313.

Isserow, J. (2018). Moral Worth and Doing the Right Thing by Accident. *Australasian Journal of Philosophy,* 97 (2), 251-264.

Jackson, F. & Pettit, P. (1988). Functionalism and broad content. *Mind,* 97, 381-400.

Jackson, F. & Pettit, P. (1990a). Program explanation: A general perspective. *Analysis,* 50, 107-117.

Jackson, F. & Pettit, P. (1990b). Causation and the philosophy of mind. *Philosophy and Phenomenological Research,* 50, 195-214.

Jackson, Frank. (1991). Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics*,101 (3), 461–482.

Jackson, F. & Pettit, P. (1992). Structural explanation in social theory. In D. Charles & K. Lermon (eds), *Reduction, Explanation, and Realism*. Oxford: Clarendon Press, 97-131.

Jackson, F. (1995). Essentialism, Mental Properties and Causation. *Proceedings of the Aristotelian Society,* 95, 253-268.

Jacobs, J. D. (2010). A powers theory of modality: or, how I learned to stop worrying and reject possible worlds. *Philosophical Studies,* 151 (2), 227-248

Jaster, R. (2020). *Agents' Abilities*. Berlin, New York: De Gruyter.

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, 68, 221–263.

Johnson King, Z. (2020). Accidentally Doing the Right Thing. *Philosophy and Phenomenological Research,* 100 (1), 186-206.

Kane, R. (1996). *The Signifi cance of Free Will*. Oxford: Oxford University Press.

Kiesewetter, B. (2017). *The Normativity of Rationality*. Oxford: Oxford University Press.

Kim, J. (1988) Explanatory Realism, Causal Realism, and Explanatory Exclusion. *Midwest Studies in Philosophy,* XII, 225-239.

Kim, J. (1990). Explanatory Exclusion and the Problem of Mental Causation. In E. Villaneuva (ed.), *Information, Semantics and Epistemology*. Blackwell, Cambridge, 36-56

Kistler, M. (2007). The Causal Efficacy of Macroscopic Dispositional Properties. In M. Kistler & B. Gnassounou (eds.), *Dispositions and Causal Powers*. Ashgate, 103-132.

Koons, R. C. (1998). Teleology as higher-order causation: A situation-theoretic account. *Minds and Machines,* 8 (4), 559-585.

Korsgaard, C. M. (1986). Skepticism about practical reason. *Journal of Philosophy,* 83 (1), 5–25.

Korsgaard, C. M. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Lando, T. (2017). Coincidence and Common Cause. *Noûs,* 51 (1), 132-151.

Lange, M. (2010). What Are Mathematical Coincidences? *Mind,* 119 (474), 307–40.

Lange, M. (2016). *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford University Press USA.

Lavin, D. (2004). Practical Reason and the Possibility of Error. *Ethics*, 114(3), 424-457.

Lehrer, K. (1968). Cans without ifs. *Analysis,* 29(1), 29-32.

Leon, F. & Tognazzini, N. A. (2010). Why Frankfurt-Examples Don't Need to Succeed to Succeed. *Philosophy and Phenomenological Research,* 80 (3), 551-565.

Lewis, D. (1973a). *Counterfactuals*. Oxford: Blackwell.

Lewis, David (1973b). Counterfactuals and Comparative Possibility. *Journal of Philosophical Logic*, 2(4), 418–446.

Lewis, D., (1976). The Paradoxes of Time Travel. *American Philosophical Quarterly,* 13, 145-152.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Nous*, 13, 455–476.

Lewis, D. (1997). Finkish Dispositions. *The Philosophical Quarterly,* 47, 143–158.

Lewis, D. (1986). *Philosophical papers: Volume II*. New York: Oxford University Press.

Lindeman, K. (2017). Constitutivism without Normative Thresholds. *Journal of Ethics and Social Philosophy*, 12(3), 231-258

Littlejohn, C. (2012). *Justification and the Truth-Connection*. Cambridge University Press.

Littlejohn, C. (2014a). The unity of reason. In C. Littlejohn & J. Turri (Eds.), *Epistemic norms*. Oxford: Oxford University Press, 135–154.

Littlejohn, C. (2014b). Fake Barns and false dilemmas. *Episteme,* 11 (4), 369-389.

Lord, E. (2007). Dancy on Acting for the Right Reason. *Journal of Ethics and Social Philosophy,* (3), 1-7.

Lord, E. (2008). Dancy on acting for the right reason. *Journal of Ethics and Social Philosophy,* 2 (3), 1–6.

Lord, E. (2010). Having reasons and the factoring account. *Philosophical Studies*, 149 (3), 283 - 296.

E. Lord & B. Maguire (eds.) (2016). *Weighing Reasons*. Oxford University Press USA.

Lord, E. (2018). *The Importance of Being Rational.* Oxford, UK: Oxford University Press.

Luper-Foy, S. (1984). The Epistemic Predicament: Knowledge, Nozickian Tracking, and Scepticism. *Australasian Journal of Philosophy,* 62 (1), 26–49.

Lutz, M. (2020). Explanationism provides the best explanation of the epistemic significance of peer disagreement. *Philosophical Studies,* 177 (7),1811-1828.

Maier, J. (2014). Abilities. *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2014/entries/abilities/>.

Maier, J. (2013). The Agentive Modalities. *Philosophy and Phenomenological Research,* 87 (3), 113-134.

Malzkorn, W. (2000). Realism, functionalism and the conditional analysis of dispositions. *Philosophical Quarterly,* 50, 452–469.

Manley, D. & Wasserman, R. (2007). A gradable approach to dispositions. *Philosophical Quarterly,* 57 (226), 68–75.

Manley, D. & Wasserman, R. (2008). On linking dispositions and conditionals. *Mind,* 117 (465), 59-84.

Mantel, S. (2014). No reason for identity: On the relation between motivating and normative reasons. *Philosophical Explorations,* 17 (1), 49–62.

Mantel, S. (2017). Three Cheers for Dispositions: A Dispositional Approach to Acting for a Normative Reason. *Erkenntnis,* 82 (3), 561-582.

Mantel, S. (2018). *Determined by Reasons: A Competence Account of Acting for a Normative Reason*. New York, USA: Routledge.

Marcus, Eric (2012). *Rational Causation*. Harvard University Press.

Markovits, J. (2010). Acting for the Right Reasons. *The Philosophical Review*, 119 (2), 201–242.

Martin, C. B. (1994). Dispositions and Conditionals. *The Philosophical Quarterly,* 44, 1–8.

Martin C. B. & Heil J. (1998). Rules and Powers. *Nous,* 12, 283-312.

Martin, C. B. (2007). *The Mind in Nature*. Oxford University Press.

Mayr, E. (2011). *Understanding Human Agency*. Oxford University Press.

McDowell, J. (1979). Virtue and Reason. *Monist*, 62, 331–350.

McDowell, J. (1995). Might There Be External Reasons? In J.E.J. Altham and R. Harrison (eds.), *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*, Cambridge: Cambridge University Press, 387–398.

McDowell, J. (2011). *Perception as a Capacity for Knowledge*. Marquette University Press.

McKenna, M. (2001). Review of John Martin Fischer and Mark Ravizza's Responsibility & Control. *Journal of Philosophy*, XCVIII, no. 2, 93–100.

McKenna, M. (2003). Robustness, Control, and the Demand for Morally Significant Alternatives: Frankfurt Examples with Oodles and Oodles of Alternatives. In Widerker D. and McKenna M. (eds.) *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. Ashgate. 201–217.

McKenna, M. (2008). Frankfurt's Argument against Alternative Possibilities: Looking beyond the Examples. *Noûs,* 42, 770–93McKenna, M. (2013). Reasons-Responsiveness, Agents, and Mechanisms. In: *Oxford Studies in Agency and Responsibility Volume 1*. Oxford University Press.

Mckitrick, J. (2005). Are Dispositions Causally Relevant? *Synthese,* 144 (3):357-371.

McKitrick, J. (2010). Manifestations as effects. In Anna Marmodoro (ed.), *The Metaphysics of Powers: Their Grounding and Their Manifestations*. Routledge.

McNaughton, D. (1988). *Moral Vision*. Oxford: Basil Blackwell.

Melden, A.I. (1961). *Free Action*, London: Routledge and Kegan Paul.

Mele, A. R. (1992a). Acting for Reasons and Acting Intentionally. *Pacific Philosophical Quarterly,* 73, 355-357

Mele, Alfred R. (1992b). *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

Mele, A. (1995). *Autonomous Agents*. New York, NY: Oxford University Press.

Mele, A.R. (2003). *Motivation and Agency*, Oxford: Oxford University Press.

Mele, A. R. (2006). *Free Will and Luck*. New York, NY: Oxford University Press.

Mele, Alfred R. (2007). Reasonology and False Beliefs. *Philosophical Papers,* 36 (1),91-118.

Mellor, D. H. (1995). *The Facts of Causation.* Routledge.

Michon, C. (2007). Opium's Virtus Dormitiva. In B. Gnassounou & M. Kistler (eds.), *Dispositions and Causal Powers*. Ashgate, 133-150.

Millar, A. (2004). *Understanding People: Normativity and Rationalizing Explanation*. Oxford University Press UK.

Millar, A. (2009). What is it that cognitive abilities are abilities to do? *Acta Analytica,* 24 (4), 223-236.

Millar, A. (2019). *Knowing by Perceiving*. Oxford University Press

Miller, C. (2008). Motivation in agents. *Noûs,* 42 (2), 222–266.

Miracchi, L. (2015). Competence to Know. *Philosophical Studies,* 172(1), 29-56.

Moliere. (1935). Le malade imaginaire, com?die-ballet ...avec une notice biographique, une notice litteraire et des notes explicatives par Ren? Vanbourdolle, Librairie Hachette, Paris.

Molnar, G. (1999). Are dispositions reducible? *Philosophical Quarterly,* 49 (194), 1-17.

Molnar, G. (2003). *Powers: A Study in Metaphysics.* Oxford University Press.

Monod, J. (1970). *Le Hasard et la Necessite: Essai sur la philosophie naturelle de la biologie moderne*. Editions du Seuil.

Moore, G. E. (1912). *Ethics*. Oxford University Press.

Morton, A. (1980). *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Oxford University Press.

Moya, C. J. (2007). Moral Responsibility Without Alternative Possibilities? *Journal of Philosophy,* 104 (9), 475-486.

Mumford, S. (1998). *Dispositions*. Oxford: Oxford University Press.

Mumford, S. (2009). Passing Powers Around. *Monist,* 92, 94-111.

Mumford S. & Anjum, R. (2011). Getting Causes from Powers. Oxford: Oxford University Press.

Nadler, S. (2010). *Occasionalism: Causation Among the Cartesians*. Oxford University Press.

Nagel, T. (1970). *The Possibility of Altruism*, Oxford: Oxford University Press.

Nagel, T. (1979). *Mortal Questions*. Cambridge University Press.

Needham, P. (1988). Causation: Relation or Connective? *Dialectica,* 42 (3), 201-220.

Nelkin, D. (2011). Making Sense of Freedom and Responsibility. Oxford: Clarendon Press.

Nolan, Daniel (2014). Hyperintensional metaphysics. *Philosophical Studies,* 171 (1),149-160.

Nozick, R. (1981). *Philosophical Explanations*. Belknap Press of Harvard University Press.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*, Oxford: Oxford University Press.

O'Connor, T. & Sandis, C. (eds.) (2010). *A Companion to the Philosophy of Action.* Wiley-Blackwell.

Otsuka, M. (1998). Incompatibilism and the avoidability of blame. *Ethics,* 108, 685-701.

Owens D. (1990). Causes and coincidences. *Proceedings of the Aristotelian Society*, 90, 49–64.

Owens D. (1992). *Causes and Coincidences*. Cambridge University Press.

Parfit, D. 2001. Rationality and Reasons. In D. Egonsson, B. Petterson and T. Ronnow-Rasmussen (eds.), *Exploring Practical Philosophy*. Aldershot: Ashgate

Parfit, D. (2011). *On What Matters: Two-Volume Set*. Oxford University Press.

Parfit, D. (1997). Reasons and motivation. *Aristotelian Society, Supplementary,* 71 (1), 99–130.

Parsons, T. (1990). *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.

Peacocke, C. (1979a). *Holistic Explanation: Action, Space, Interpretation*. Clarendon Press.

Peacocke, C. (1979b). Deviant Causal Chains. *Midwest Studies in Philosophy* 4 (1):123-155.

Pears, D. (1975). The Appropiate Causation of Intentional Basic Actions. *Critica* 7 (20), 39-72.

Pereboom, D. (1995). Determinism al Dente. *Noûs,* 29, 21–45.

Pereboom, D. (2000). Alternate Possibilities and Causal Histories. *Philosophical Perspectives*, 14, 119-138.

Perboom, D (2001). *Living Without Free Will*, Cambridge: Cambridge University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Pettit, P. and Smith, M. (1990). Backgrounding Desire. *Philosophical Review,* 99, 565-592.

Pettit, P and Smith, M. (1993). Practical Unreason. *Mind,* 102, 53-79.

Pettit, P and Smith, M. (1996). Freedom in Belief and Desire. *Journal of Philosophy,* 93, 429-449.

Pettit, P and Smith, M (2006). External Reasons. In C. Macdonald and G. Macdonald (eds), *McDowell and His Critics.* Oxford: Blackwell.

Pollock, J. & Cruz, J. (1999). *Contemporary Theories of Knowledge, 2nd Edition*. Rowman & Littlefield.

Platts M. (1980). Moral Reality and the End of Desire. In M. Platts (ed.), *Reference, Truth, and Reality*, London: Routledge and Kegan Paul.

Prichard, H.A. (1932). Duty and Ignorance of Fact' Reprinted in *Moral Writings*. Oxford: Clarendon Press (2002).

Prior, E. W, Pargetter, R. and Jackson, F. (1982). Three Theses about Dispositions. *American Philosophical Quarterly,* 19(3), 251-257.

Prior, E. W. (1985). *Dispositions.* Aberdeen: Aberdeen University Press.

Pritchard, D. (2005). *Epistemic Luck*. Oxford: Oxford University Press.

Pritchard, D. (2007). Anti-Luck Epistemology. *Synthese,* 158 (3), 277–97.

Pritchard, D. (2008). Knowledge, luck and lotteries. In Hendricks, V. (ed), *New Waves in Epistemology*. Palgrave Macmillan.

Pritchard D. (2009). Safety-Based Epistemology: Whither Now? *Journal of Philosophical Research,* 34, 33–45.

Pritchard, D. (2012). Anti-Luck Virtue Epistemology. *Journal of Philosophy,* 109 (3), 247-279.

Pritchard, D. & Whittington, L. J. (eds.) (2015). *The Philosophy of Luck*. Wiley-Blackwell.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge University Press.

Quinn, W. (1993). Putting Rationality in its Place. In *Morality and Action*. Cambridge University Press, 228-255

Raz, J. (1975). *Practical Reasoning and Norms*, London: Hutchinson & Co., reprinted, Oxford University Press.

Raz, J. (1999). *Engaging Reason: On the Theory of Value and Action*, Oxford: Oxford University Press

Raz, J. (2005). The Myth of Instrumental Rationality. *Journal of Ethics and Social Philosophy*, 1 (1), 28.

Raven, M. J. (2015). Ground. *Philosophy Compass,* 10 (5), 322-333.

Reutlinger, A. (2016). Is There A Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science, 83* (5), 733-745.

Riggs, W. D. (2014). Luck, Knowledge, and "Mere" Coincidence. *Metaphilosophy,* 45 (4-5), 627-639.

Ross, W. D. (1939). *Foundations of Ethics*. Oxford: Clarendon Press.

Roush, S. (2005). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford University Press.

Ruben, D.-H. (1994). A counterfactual theory of causal explanation. *Nous*, 28 (4), 465-4

Russell, B. (1948). *Human Knowledge: Its Scope and Limits*. London: Allen & Unwin.

Ryle, G. (1949). *Filling in Space. The Concept of Mind*. Cambridge: Cambridge University Press

Sainsbury, R. M. (1997). Easy Possibilities. *Philosophy and Phenomenological Research,* 57, 907–19.

Sandis, C. (2006). The Explanation of Action in History. *Essays in Philosophy* 7, (2),12.

Sartorio, C. (2015). Sensitivity to Reasons and Actual Sequences. In Shoemaker, D. (ed.*), Oxford Studies in Agency and Responsibility*, Volume 3. Oxford University Press UK.

Sartorio, C. (2016b). Vihvelin on Frankfurt-Style Cases and the Actual-Sequence View. *Criminal Law and Philosophy,* 10 (4), 875-888.

Sartorio, C. (2016a). *Causation and Free Will*. Oxford University Press UK.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Scanlon, T. M. (2008). *Moral Dimensions. Permissibility, Meaning, Blame*. Cambridge, MA: Belknap Press.

Scanlon, T. M. (2014). *Being Realistic About Reasons*. Oxford University Press.

Schaffer, J. (2017). The Ground Between the Gaps. *Philosophers' Imprint,* 17.

Schlosser, M. E. (2007). Basic deviance reconsidered. *Analysis,* 67 (3),186–194.

Schlosser, M. E. (2011). The Metaphysics of Rule-Following. *Philosophical Studies,* 155 (3), 345-369.

Schnall, I. M. (2010). Weak reasons-responsiveness meets its match: in defense of David Widerker's attack on PAP. *Philosophical Studies,* 150 (2), 271 – 283

Schroeder, M. (2007). *Slaves of the Passions*. Oxford: Oxford University Press

Schroeder, M. (2008). Having reasons. *Philosophical Studies*, 739(1), 57-71

Schroeder, Mark (2009). Means-end coherence, stringency, and subjective reasons. *Philosophical Studies,* 143 (2), 223 - 248.

Sehon, S. (2005). *Teleological Realism: Mind, Agency, and Explanation*, Cambridge, MA: MIT Press.

Sehon, S. (2016). *Free Will and Action Explanation: A Non-Causal, Compatibilist Account*. Oxford University Press UK.

Setiya, K. (2007). *Reasons Without Rationalism*. Princeton University Press.

Shafer-Landau, R. (2003). *Moral Realism: A Defence*. Oxford: Clarendon Press

Shope, R. K. (1992). You know what you falsely believe (or: Pollock, know thyself!). *Philosophy and Phenomenological Research,* 52 (2), 405-410.

Singh, K. (2019). Acting and Believing Under the Guise of Normative Reasons. *Philosophy and Phenomenological Research,* 99 (2), 409-430.

Skorupski, J. (2010). *The Domain of Reasons*. Oxford: Oxford University Press

Skow, B. (2016). *Reasons Why*. Oxford University Press UK.

Sliwa, P. (2016). Moral Worth and Moral Knowledge. *Philosophy and Phenomenological Research, 93*(2), 393–418

Smart, J. J. C. (1961). Free-will, praise and blame. *Mind*, 70(279), 291-306

Smith, A. D. (1977). Dispositional Properties. *Mind*, 86, 439–445.

Smith, M. (1994). *The Moral Problem*. Blackwell.

Smith, M. (1997). A theory of freedom and responsibility. In Cullity, G. & Gaut, B. (eds.), *Ethics and Practical Reason*. Oxford University Press, 293-317.

Smith, M. (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In Stroud, S. & Tappolet, C. (eds.), *Weakness of Will and Practical Irrationality*. Oxford: Clarendon Press, 17-38.

Smith, M. (2004). The Structure of Orthonomy. *Royal Institute of Philosophy Supplement,* 55, 165-193.

Smith, M. (2007). Is there a nexus between reasons and rationality? *Poznan Studies in the Philosophy of the Sciences and the Humanities,* 94 (1), 279-298.

Smith, M. (2009). The explanatory role of being rational. In Sobel, D. & Wall, S. (eds.), *Reasons for Action*. Cambridge: Cambridge University Press, 58-80.

Sober E. (1984). Common cause explanation. *Philosophy of Science*.

Sorensen, K. (2014). Counterfactual Situations and Moral Worth. *Journal of Moral Philosophy,* 11 (3), 294-319.

Sosa, E. (1999a). How Must Knowledge Be Modally Related to What Is Known? *Philosophical Topics,* 26 (1/2), 373–384.

Sosa, E. (1999b). How to Defeat Opposition to Moore. *Noûs,* 33 (s13), 141–53.

Sosa, E. (2002). Tracking, Competence, and Knowledge. In Moser, P. (ed.), *The Oxford Handbook of Epistemology*. Oxford University Press, Oxford.

Sosa E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. *Vol. I*. Oxford University Press.

Sosa E (2009). *Reflective Knowledge: Apt Belief and Reflective Knowledge*. *Vol. II*. Oxford University Press.

Sosa, E. (2010). *Knowing Full Well*. Princeton University Press.

Sosa, E. (2017). *Epistemology*. Princeton: Princeton University Press.

Spencer, J. (2017). Able to Do the Impossible. *Mind,* 126 (502), 466-497.

Stalnaker, R. (1968). A theory of conditionals. *Studies in Logical Theory, American Philosophical Quarterly Monograph Series*, 2. Oxford: Blackwell, 98–112.

Stampe, D. W. (1977). Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy,* 2 (1), 42-63.

Steward, H. (1997). *The Ontology of Mind: Events, Processes, and States*. Oxford University Press.

Steward, H. (2009). Fairness, Agency and the Flicker of Freedom. *Noûs,* 43 (1), 64 - 93.

Steward, H. (2008). Moral responsibility and the irrelevance of physics: Fischer's semi-compatibilism vs. anti-fundamentalism. *Journal of Ethics*, 12, 129-145.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford University Press.

Stoecker, R. (1993). Reasons, Actions, and their Relationship. In Stecker (ed), *Donald Davidson Responding to an International Forum of Philosophers*. Berlin and New York De Gruyter.

Stoecker, R. (2003). Climbers, Pigs and Wiggled Ears – The Problem of Waywardness in Action Theory. In Walter, S. and Heckmann, H. D. (eds.), *Physicalism and Mental Causation*. Imprint Academic.

Stoecker, R. (2001). Agents in action. *Grazer Philosophische Studien*, 61, 21.

Stout, R. (1996). *Things That Happen Because They Should: A Teleological Approach to Action*. Oxford University Press.

Stout, R. (2005). *Action*. Routledge.

Stout, R. (2010). Deviant Causal Chains. In O'Connor and Sandis (2010), 159-156.

Stoutland, F. 2001. "Responsive action and the belief-desire model." *Grazer Philosophische Studien* 61 (1): 83–106

Stratton-Lake, P. (2000). *Kant, Duty and Moral Worth*. London: Routledge

Strawson, P. F. (1985). Causation and explanation. In Vermazen, B. & Hintikka, M.B. (eds.), *Essays on Davidson: Actions and Events*. Oxford University Press, 115--35.

Stump, E. (1996). Libertarian Freedom and the Principle of Alternative Possibilities. In Howard-Snyder, D. and Jordan, J. (eds), *Faith, Freedom, and Rationality: Essays in the Philosophy of Religion*. Lanham, MD: Rowman and Littlefield.

Stump, E. (2003). Moral responsibility without alternative possibilities. In Widerker, D. & McKenna, M. (eds.), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. Ashgate, 139-158.

Sylvan, K. (2015). What apparent reasons appear to be. *Philosophical Studies*, 172 (3), 587-606.

Sylvan, K. & Lord, E. (2019). Prime Time (for the Basing Relation). In Carter, J.A. & Bondy, P. (eds.), *Well-Founded Belief: New Essays on the Basing Relation*.

Todd, P. & Tognazzini, N. A. (2008). A problem for guidance control. *Philosophical Quarterly,* 58 (233), 685-692.

Turri, J. (2009). The ontology of epistemic reasons. *Noûs*,43, 490–512

Turri, J. (2011). Manifest Failure: The Gettier Problem Solved. *Philosophers' Imprint,* 11.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon.

Velleman, D. (1992a). What Happens When Someone Acts? *Mind,* 101 (403), 461-481.

Velleman, D. (1992b). The Guise of the Good. *Nous*, 26, 3-26.

Vetter, B. (2013). Multi-track dispositions. *Philosophical Quarterly,* 63 (251), 330-352.

Vetter, B. (2015). *Potentiality: From Dispositions to Modality*. Oxford University Press.

Vetter, B. & Jaster, R. (2017). Dispositional accounts of abilities. *Philosophy Compass,* 12 (8), 12432.

Vetter, B. (forthcoming). Are abilities dispositions? *Synthese,* 196 (1).

Vihvelin, K. (2000). Freedom, Foreknowledge, and the Principle of Alternate Possibilities. *Canadian Journal of Philosophy,* 30 (1), 1-23.

Vihvelin, K. (2004). Free Will Demystified: A Dispositional Account. *Philosophical Topics, 32*, 427–450.

Vihvelin, K. (2008). Foreknowledge, Frankfurt, and ability to do otherwise: A reply to Fischer. *Canadian Journal of Philosophy,* 38 (3), 343-372.

Vihvelin, K. (2013). *Causes, Laws, & Free Will*. New York: Oxford University Press.

von Wright G. H. (1963). *Norm and Action: A Logical Enquiry*. Humanities Press, New York, 35-36.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.

Watson, G. (1975). Free Agency. Reprinted in Watson (1982).

Watson, G. (1977). Skepticism about Weakness of Will. *Philosophical Review,* 86, 316-39.

Watson G. (ed.), (1982). *Free Will*. New York: Oxford University Press.

Watson, G. (2004). *Agency and Answerability*. New York: Oxford University Press.

Way, J. (2009). Two Accounts of the Normativity of Rationality. *JESP*, 4.2, 1–8.

Wedgwood, R. (2004). The metaethicists' mistake. *Philosophical Perspectives,* 18 (1), 405–426.

Wedgwood, R. (2006). The normative force of reasoning. *Noûs,* 40 (4), 660–686.

Weslake, B. (2010). Explanatory depth. *Philosophy of Science, 77* (2), 273–294.

Whiting, D. (2014). Keep Things in Perspective: Reasons, Rationality, and the A Priori. *Journal of Ethics and Social Philosophy,* 8 (1), 1-22.

Whittle, A. (2010). Dispositional Abilities. *Philosophers' Imprint*, 10(12) (September), 1-23.

Wiggins, D., (1987). A Sensible Subjectivism? In *Needs, Values, and Truth*. Oxford: Blackwell, pp. 185–214.

Williams, B. (1979). Internal and external reasons. Reprinted in B. Williams (1981), *Moral luck*. Cambridge: Cambridge University Press, 101–113.

Williams, N. E. (2019). *The Powers Metaphysic*. Oxford University Press.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Wilson, R. A. (1994). Causal Depth, Theoretical Appropriateness, and Individualism in Psychology. *Philosophy of Science, 61*(1), 55–75.

Wolf, S. (1990). *Freedom Within Reason*. Oup Usa.

Woodward, J. (2003). *Making things happen*. New York: Oxford University Press

Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In Thagard, P. (Ed.), *Philosophy of psychology and cognitive science.* New York: Elsevier, 31–79.

Wright, C. (2012). Mechanistic explanation without the ontic conception. *European Journal for Philosophy of Science*.

Wu, W. (2016). Experts and Deviants: The Story of Agentive Control. *Philosophy and Phenomenological Research*, 92(2), 101–26.

Wyma, K. (1997). Moral responsibility and leeway for action. *American Philosophical Quarterly,* 34, 57–70.

Yablo, S. (1992). Mental Causation. *The Philosophical Review, 101*(2), 245–280.

Zagzebski, L. (2000). Does Libertarian Freedom Require Alternate Possibilities? *Philosophical Perspectives,* 14, 231–48.

Zimmerman, M. J. (1987). Luck and moral responsibility. *Ethics*, 97(2), 374–386.

Zimmerman, Michael J. (2002). Taking luck seriously. *Journal of Philosophy,* 99 (11), 553-576.