The Darker Steed

Reason, Passion and Self-Awareness

Edgar Haydon Phillips

UCL

A thesis submitted for the degree of
Ph.D. in Philosophy

I, Edgar Haydon Phillips, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

.....

Chapters 1 and 2 are adapted, with substantial changes, from material that first appeared in my MPhil Stud thesis, *From the Agent's Point of View* (2015).

Abstract

Reasons 'favour' and justify actions, but they also explain our actions. Because we are self-aware, rational agents whose actions are guided by our appreciation of what reasons we have to act, these explanatory and justificatory roles are not wholly separate. A person's reasons for acting make sense of their action from their point of view as its agent: they show us why the person did what they did by showing us what point they saw in doing it. There is, however, a tension within the idea of reasons as normative and explanatory. Considered as normative, it is natural to think of reasons as objective and universal: reasons are backed up by normative principles, and if something is a reason for me to act in a certain way, it would be a reason for anyone in relevantly similar circumstances to do the same. But explaining a person's actions from their point of view—showing the point they saw in doing what they did—often introduces elements of idiosyncrasy, in particular when an action is explained by false beliefs or quirky desires.

Belief's role, I argue, is easily accommodated by the universalistic conception. Reasons are facts; because we make mistakes about the facts, we can make mistakes about our reasons. In these cases, understanding my action from my perspective simply requires an appreciation of my perspective on what universal reasons I had. Desire, however, poses a more serious challenge. Many desires cannot be understood just by considering their subject's perspective on universal reasons, but they can and do figure ineliminably in our understanding of our own actions. We thus need to recognise that some reasons are not universal but irreducibly personal and particular. There is thus a plurality within reasons for action: reason is universal, and it is idiosyncratic.

Impact Statement

Contemporary philosophical accounts of reasons for action have tended to embrace one of two extremes: our reasons for action are either taken to be wholly universal and objective, based in universal values, principles or judgements; or they are taken to be entirely subjective and idiosyncratic, based in desires and motivations that are simply 'given'. Each extreme fails to accommodate what truth there is in the other. This thesis makes a case for this idea—that there is some truth in each view, because there is a plurality in the sources of our reasons. If its arguments are accepted, this might encourage further investigation of some under-explored but potentially fertile ground.

The fourth chapter of the thesis engages with work in developmental psychology and criticises widely-held theoretical assumptions about how we should understand so-called 'theory of mind', namely that a fully-developed view of the mind understands all mental states in representational terms, as 'propositional attitudes'. I hope to publish a version of this chapter in an interdisciplinary journal for philosophy of mind and psychology. This will hopefully influence both philosophers of mind and researchers in psychology to take more seriously the diversity and complexity of our mature understanding of the mind.

Contents

Acknowledgements

It is an extraordinary privilege to have the opportunity to undertake a PhD in philosophy, no less to be given the resources and the support necessary to complete it. I am grateful to the AHRC and the London Arts and Humanities Partnership for funding me from 2015–18, and to UCL and in particular the department of philosophy for providing the resources and an environment that enabled me to do this work. I should also acknowledge the support of the UCL Old Students' Association, who helped to fund the second year of my MPhil Stud, from which parts of this thesis developed.

As well as material resources, completing a PhD, especially in philosophy, requires a very different kind of support—the intellectual engagement, guidance, and emotional support of one's peers and teachers. While the process of writing a thesis can feel very lonely at times, it is absolutely not something that can be done on one's own. This thesis would not exist without the intelligence, kindness, criticism, and friendship of a great many people.

First and foremost among these is Mike Martin. I hope Mike will not mind me saying that he can be quite a formidable figure. His reputation as a tenacious inquisitor precedes him, but I cannot overstate how glad I am that I eventually took the risk (as it seemed at the time) of asking to work with him. As a student of Mike's, one quickly learns that he is not only a tireless opponent of lazy thinking but also an exceptionally patient, generous and encouraging supervisor. I count myself extremely fortunate to be among those who have benefited enormously from each of these aspects of his unique philosophical character, and both I and this thesis owe more to his guidance than I can express. I must also thank Mike specifically for getting me to see the truth of the central idea that this thesis attempts to motivate, and for pointing me to the central example it uses to do so.

I would also especially like to thank Ulrike Heuer and Lucy O'Brien. Ulrike has been my main source of supervisory support at UCL for the last year, and approached what was already a fairly developed project with exactly the balance of sympathy, criticism and enthusiasm that was needed. Lucy supervised me for most of the MPhil Stud and for the start of the PhD. She helped me find my voice when I was still quite new to philosophy. I am forever grateful to them both. I have also benefited from the wisdom and support of many others whilst studying in London—more than I could name here. I do, however, want specifically to thank: Amia Srinivasan and Doug Lavin for insightful comments and helpful discussion on parts of this thesis; Maria Alvarez, Clayton Littlejohn and John Hyman for helping, in different ways, to shape my philosophical interests; Fiona Leigh for her invaluable pastoral support and for teaching me about Plato and Aristotle; and Richard Edwards for years of tireless work helping with all aspects of admin. I am also grateful to everyone else who taught me during my MPhil Stud at UCL and my MA at King's.

Working in London, and specifically at UCL, I have been part of a truly special graduate community, and I have been tremendously lucky to have so many brilliant and lovely philosophical friends, including Julian Bacharach, Showkat Ali, Ilaria Cozzaglio, David Olbrich, Vanessa Carr, Alec Hinshelwood, Polly Mitchell, Andrew Knox, Bárbara Núñez de

Cáceres, Catherine Dale, Michael Markunas, James Laing, Niels Christensen, Shunichi Takagi, Alex Geddes, Pete Faulconbridge, Henry Clarke, Léa Salje, Ashley Shaw and Paul Doody. I apologise to those I have failed to mention by name. I want to give special thanks to Mog Hampson, Jerome Pedro, Alex Sayegh, Tim Short, Tom Williams and Karine Sawan for letting me in their gang, or at least letting me hang out with their gang. And I want to give very, very special thanks to Charles Jansen and Laura Silva for being better friends than I ever thought I would have.

More than anything else, this thesis owes its existence to Karoline Phillips, without whose love, support and impatient, no-nonsense encouragement I could not have written it, and whose faith in me allowed me to feel some faith in myself. Thank you, Karoline.

Finally, I want to thank Hilary and Bill Phillips. I have been extremely lucky to have such kind, generous and loving parents, and I dedicate this thesis to them.

Now, the thing was that Hans Castorp, for a long time, had had his eye upon this Pribislav; had chosen him out of the whole host, known and unknown, in the court-yard of the school, taken an interest in him, followed him with his eyes – shall we say admired him? – at all events observed him with peculiar sympathy. Even on the way to school he looked forward with pleasure to watching him among his fellows, seeing him speak and laugh, singling out his voice from the others by its pleasantly veiled, husky quality. Granted that there was no sufficient ground for his preference, unless one might refer it to Hippe's heathenish name, his character as model pupil – this latter was, of course, out of the question – or to the 'Kirghiz' eyes, whose grey-blue glance could sometimes melt into a mystery of darkness when one caught it musing sidewise; whichever it might be, or none of these, Hans Castorp troubled not a whit to justify his feelings, or even to question by what name they might suitably be called. … Hans Castorp was penetrated by the unconscious conviction that an inward good of this sort was above all to be guarded from definition and classification.

Thomas Mann, *The Magic Mountain*, 1924 (trans. H. T. Lowe-Porter, 1928)

Introduction

The contemporary notion of 'reasons for action' unites a number of philosophical concerns. There are questions of the good and the right, of what we ought to do, of what is important and what actions are justified. There questions about practical reasoning and rationality, how we should and do think about what to do. And there are questions about psychological explanation—why people do the things they do; what, on a given occasion, a particular person's reasons were for acting as they did. These diverse concerns are united in the rational self-awareness of human agency. When we act rationally, we aim to act well, guided by good reasons. Because we are self-aware, we understand our own actions in terms of the reasons for which we act. Because we understand others as rational, self-aware agents like ourselves, our understanding of other people's actions reflects their own understanding of their own actions. While the aspects of normativity, rationality and self-understanding can for some investigative purposes be teased apart, they are thoroughly intertwined in moral and rational psychology, in the idea of acting for a reason.

This thesis is concerned in the first instance with the way reasons explain actions. The starting point for the investigation will, in general, be a question about how we understand ourselves. Because of the interconnections just noted, though, the claims and arguments made herein will not be entirely neutral on questions of justification and rationality, of what is a good reason and what it is rational for a person to do. Indeed, the central theme will be a tension that can be seen to arise from the need to make sense of the connections between the demands of justification, rationalisation and understanding. On one hand, we are inclined to think of reasons as *universal*, in the sense that if I judge that R is a reason for me to V in the circumstances I am in, I am committed to the judgement that R would be a reason for anyone in relevantly similar circumstances to V. On the other hand, we recognise certain sources of *idiosyncrasy* in human action and acknowledge that understanding an action from the point of view of its agent often requires a recognition of such idiosyncrasy. In this thesis, I consider two kinds of idiosyncrasy and investigate how they interact with the conception of reasons for action as universal.

The first kind of idiosyncrasy is cognitive. Agents can form, and act on, beliefs that may not accurately represent how things really are. When they do so, we may not be able to understand their doing what they do just on the basis of considerations that we recognise as reasons for them to do what they did: they may do something that, as we see it, they had no reason to do. While there are difficult questions about the details of how exactly this kind of idiosyncrasy is to be best integrated with the universalistic conception of reasons, it does not, as we will see, present a very serious challenge to that conception. When an action is taken on the basis of false belief, we understand the action by coming to appreciate how, to put it roughly, the agent took themselves to have some reason of the universal kind. When we turn our attention to the second source of idiosyncrasy, however, things are not so

straightforward. States of the soul that we might broadly speaking identify as 'passions' seem to play an ineliminable role in explaining certain actions.

The idea of an opposition between reason and the passions is a very old one. In Plato's famous metaphor, the soul is 'the union of powers in a team of winged steeds and their winged charioteer' (*Phaedrus* 246a).[1] The job of reason, represented by the charioteer, is to keep the passions—the horses—in line and thus to steer the chariot in the right direction. One horse, representing *thumos* or spirit, is white, noble and good, 'a lover of glory, but with temperance and modesty' (*Phaedrus* 253d). The other is

> of crooked frame ... with thick short neck, snub nose, black skin and green eyes; hot-blooded, consorting with wantonness and vainglory; shaggy of ear, deaf, and hard to control with whip and goad. (*Phaedrus* 253e)

This abject creature represents *eros* or appetite, the kind of psychical force most naturally referred to, in modern non-technical English, as 'desire'. Desire, in this Platonic picture, is a force of corruption, something only to be restrained and subjugated—sometimes violently—by the higher faculty of reason.

The image in the *Phaedrus* of the relation between reason and passion is extreme. An account that is somewhat less openly hostile to desire can be found, for instance, in Aristotle. Desire can still, in Aristotle's picture, conflict with reason, and notably does so in both the continent and the incontinent agent. However, the part of the soul that is characterised by desire need not be violently dominated but is capable of 'sharing in' reason, 'inasmuch as it heeds it and is apt to be obedient to its commands' (*NE* I.13 1102b30).[2] In the virtuous agent, the rational part of the soul and the part characterised by desire work in harmony: the agent's desires align with what reason determines to be good and hence they are not unruly or disruptive, but are in themselves virtuous. Even on this Aristotelian view, though, desires seem to be subordinate to reason. The passions are disruptive *except* insofar as they are respond to, or at least are in agreement with, reason.

Something like this idea finds its modern expression in the view that a desire itself does not give the desirer any reason to pursue its object, that desires are based on or responsive to non-desire-given reasons, and that any passions an agent undergoes that are not in line with her assessment of her reasons are unintelligible and experienced as 'alien' or as 'mere urges'. Versions of this idea appear, for example, in the work of Maria Alvarez, Jonathan Dancy, John McDowell, Thomas Nagel, Derek Parfit, Warren Quinn, Joseph Raz, T. M. Scanlon and Gary Watson. For reasons that will hopefully become clearer later on, I believe that a version of it also appears in the work of authors, such as Simon Blackburn, Allan Gibbard and Mark Schroeder, who in a way privilege desire or desire-like states of mind in their accounts of practical thought, but who understand such states as doing their main work from, so to speak, behind the scenes.

The central claim of this thesis is that to make sense of the way in which idiosyncratic desires figure in our self-understanding, we must appreciate them as giving us reasons of a

[1]. All quotations of the *Phaedrus* are from (Plato, 1952).
[2]. (Aristotle, 2011).

distinctive kind: reasons that are personal and particular rather than public and universal. I thus reject the idea that desire is in opposition to reason except insofar as it follows reason's lead. To understand the ways in which we understand ourselves, we need to recognise that it is sometimes reasonable to allow passion a free rein. Desire, as much as reason, belongs to the human soul. It is not always a mistake to let the black horse lead the way.

Chapter 1 begins by characterising a special kind of explanation of action which I call, following well-established usage, *rationalisation*. A rationalisation not only explains an action, but it does so in a way that reveals to us the agent's own understanding of their action. A successful rationalisation enables us to see what point the agent saw in acting as they did. It is natural to say that rationalisations explain actions in terms of the agent's reasons for acting. However, this generates a tension with another natural thought, which is that reasons for action are 'worldly': they are facts that show a potential course of action to be in some respect desirable or worth taking. The tension arises from the possibility of idiosyncrasy: some actions are not rationalisable simply by stating such facts, and need to be understood in relation to something about the agent's state of mind. If we rationalise actions by citing an agent's reasons, such idiosyncrasy might seem to introduce a distinct source of reasons. This would raise a puzzle about how the two kinds of reasons are connected.

To get clearer about this issue, I distinguish 'worldly rationalisation', in which an action is explained simply by the obtaining of a worldly reason, from 'perspectival rationalisation', in which an action is explained by a fact about what the agent believed or how things seemed to them. While they cite very different facts to explain actions, these two forms of rationalisation seem to be intimately connected. The content of the agent's belief in a perspectival rationalisation typically corresponds to a consideration that might, if true, have constituted a worldly reason for the agent to act as they did. This, I argue, enables us to resolve the apparent tension between the universality of reasons and the idiosyncrasy introduced by perspectival rationalisations. Perspectival rationalisations turn out to depend on our understanding of worldly reasons: we understand the agent's action in virtue of seeing how, had things been as they took them to be, there would have been a reason to do what they did.

While Chapter 1 concerns the relation between perspectival rationalisations and worldly reasons, Chapter 2 addresses itself to the connection between perspectival and worldly rationalisations. There is a tendency to assume that the latter are in some sense reducible to the former. This idea is supported by the apparent asymmetrical dependence of worldly on perspectival rationalisations. If I went to the shops because we had run out of milk, this seems to imply that I went to the shops because I thought we had run out of milk; however, I can go to the shops because I think we have run out of milk even if we have not run out of milk, and in this case I cannot be going to the shops because we have run out of milk. The non-worldly rationalisation seems to be better proportioned to what it explains. This motivates a presumption in favour of the view that worldly rationalisations are not fundamental. However, I argue that there is good reason to accept the irreducibility of

worldly rationalisations at least in the specific case of the rationalisation of perceptually-based beliefs, since doing so makes possible an attractive account of how those beliefs are justified. It is not immediately clear whether we can generalise this argument to apply to the rationalisation of action. One way of doing so might be to appeal to something perceptual or quasi-perceptual in the rationalisation of action, such as Scanlon's idea that desire involves 'seeing' considerations as reasons. But is that idea plausible?

Chapter 3 introduces a challenge to Scanlon's conception of desire, and indeed to the broader universalistic or 'cognitivist' conception of reasons and rationalisation that Scanlon's account epitomises. The challenge is illustrated with an example, originating with Stuart Hampshire, in which an agent who is well aware of the fact that a particular option is disfavoured by the balance of worldly reasons quite intelligibly chooses that option because he has a desire for it. I consider several ways in which the cognitivist might attempt to accommodate the example, and argue that none of them is wholly satisfactory. The cognitivist responses either make the agent out to be irrational in a way that he seems not to be, or mischaracterise the nature of the reason on which he acts. I suggest that Hampshire himself gives the best account of the example: in the example, the agent's reason is his desire. The rest of the thesis seeks to clarify and defend this idea, and to explain how this desire-based reason differs from the universal reasons discussed in Chapters 1 and 2.

To understand the nature of the agent's reason in this kind of case, we need to understand the idiosyncrasy of desire. Chapter 1 gave an account of the idiosyncrasy of belief that allowed us to say that beliefs do not themselves provide reasons for action. The basic idea was that beliefs merely *represent* such reasons. Understanding actions on the basis of the agent's beliefs thus involves 'metarepresentation': thinking about the agent's representational mental states as such. A natural cognitivist approach to the idiosyncrasy of desire is to propose that it too is representational. If this is true, then understanding idiosyncratic desires ought to require metarepresentation. To investigate this hypothesis, Chapter 4 discusses the developmental psychology of mental state concepts. A good deal of evidence suggests that children only become capable of metarepresentational thinking at around age four, when they begin to pass the so-called direct false belief test. However, even much younger children seem to show an appreciation of the idiosyncrasy of desire. This suggests that our fundamental understanding of the idiosyncrasy of desire is not metarepresentational. I consider both empirical evidence and philosophical arguments that might be thought to threaten this idea and argue that they are not convincing. In doing so, I provide an argument against the view that desires must be representational because they are 'propositional attitudes'. There is good reason to believe that the kind of desire relevant for our discussion does not take a proposition as its object. Such desires, I suggest, can be understood as basically relational. However, understanding the true nature of their idiosyncrasy involves appreciating the way in which the desire-relation is grounded in the subject, rather than its object.

Chapter 5 further investigates this idea—the idea that desires can be grounded in the subject. A common form of argument is taken to show that the desirer must understand

their desire as grounded in something else, in particular in worldly reasons. Unless the desirer sees their desire as being based on such reasons, the argument goes, their desire will be unintelligible to them, or they will experience it as an alien force. Such a desire, clearly, is not apt to provide the kind of reason it was claimed to provide in the discussion of Hampshire's example. So the proponent of desire-based reasons needs to explain why the argument fails. The argument is commonly taken to demonstrate the need for a positive account of when a subject 'identifies' with her desire, rather than being alienated from it. I argue that this is a mistake: we can treat alienation as the marked or positively characterised notion, and identification as in a sense the default case. The fact that we can be alienated from our desires does not therefore show that a desire must be backed up by some other normative consideration in order for it to rationalise action.

In the course of this discussion, I distinguish the argument about identification and alienation from an argument, due to G. E. M. Anscombe, with which it is sometimes conflated. Anscombe argues that whatever is wanted must be wanted under the aspect of some 'desirability characterisation'. This argument is, I think, compelling. This might seem to support the same conclusion as the argument from alienation: a desire has to be based on an apparent worldly reason or else it is unintelligible; but if it is based on an apparent worldly reason, the desire itself need not be seen as a source of reasons. I argue that Anscombe's conclusion is in fact consistent with the view that some reasons are based in desires, because we can understand desire itself as providing a desirability characterisation. If we distinguish Anscombe's 'wanting' from our 'desire', this turns out not to be as strange an idea as it might at first sound.

The final chapter seeks to further explore the personal and particular character of desire-based reasons in a somewhat indirect way, by considering another phenomenon that seems to share these features, namely love. The particularity of love is brought out by arguments against the idea that love is rationally based on qualities of the beloved. One of the most forceful exponents of these arguments, Niko Kolodny, rightly holds that they suggest that love is in fact relational. However, Kolodny also argues that this shows that the reason for love is the fact that one has a valuable relationship with the beloved. Kolodny's account might then seem to offer a cognitivist explanation of love's particularity. I argue that Kolodny's relationship view, precisely because of its cognitivist character, leads to implausible claims about when love should and should not make sense from the subject's perspective. I suggest that Kolodny's mistake lies in reducing the psychological role of history in understanding love to the rational role of knowledge of a certain fact. That history in fact provides a different kind of understanding from universal reasons, and this corresponds to the personal nature of love and the reasons it provides. I conclude with some reflections on the implications of this discussion for our picture of rational self-understanding.

Chapter 1

The Agent's Point of View

1.1 The varieties of rationalisation

Like events of any other kind, human actions can, given the right context, be explained by a tremendously diverse range of considerations. Some of the more common forms of explanation of actions appeal to: the agent's character; the specific circumstances of the action; the agent's needs, ends or interests, or those of someone else or of some group; their personal 'values' and ethical beliefs; their general mental state at the time of acting (for example if they were tired, excited or distracted); their desires, conscious or unconscious; their biases or prejudices; their culture or upbringing; the conditions of their childhood (whether privileged or deprived, for instance); their hopes and fears; their abilities, including intellectual abilities; their self-image; their profession or other social role; what is right or required morally, legally, or by some other system of norms; some benefit or good that might be achieved by their acting as they did.

Some of these forms of explanation seem to be more central than others to our understanding of human action as such. In particular there seems to be an important connection between acting intentionally—which we might regard as the paradigm of human action—and acting for a reason.[3] When we act intentionally, deviant cases aside, we act knowing what we are doing and knowing why we are doing it.[4] In Anscombe's formulation, an intentional action is one to which a certain 'Why?'-question has application, that being the 'Why?'-question that asks for a reason for the action. An answer to this question, if positive, reveals what *point* the agent sees in doing what they are doing.[5] This is, in the first instance, a question addressed to the agent themselves, and it is, given the self-awareness of

[3]. I do not mean to say that it is obvious that every intentional action is done for a reason. For different developments of the idea that there is a connection here, see in particular (Anscombe, 1963 who does not take every intentional action to be done for a reason; Davidson, 1980a who apparently does).

[4]. Again see (Anscombe, 1963). Strictly this claim needs to be qualified to say that we know what we are doing 'under a description'—the description under which the action is intentional. There is also usually (perhaps always) a range of descriptions of any action of which the agent is ignorant; these descriptions, though, do not tell us what the person is 'up to', and the action is not rationalised under such descriptions.

intentional human action, a question to which the agent should normally have an answer. Knowing what I am up to in doing what I am doing, I know why I am doing what I am doing here and know. I am typing this sentence, for example, in order to make a point about self-awareness in intentional action; I am typing it because by doing so I might convey something important about the first-person perspective on action and reasons for action.

Certain forms of explanation of action, more than others, reveal this first-person perspective: they explain action, as we might say, from the agent's point of view. When an action is explained to us in this way, the understanding we acquire of the agent's behaviour mirrors, in a way, the agent's understanding of themselves. Because in central cases the agent's understanding of why they are doing what they are doing is intimately tied up with their reasoning about what *to* do, with reasons *for* doing what they are doing, this kind of first-personal explanation characteristically works by telling us how, from the agent's perspective, their action was, in some sense, a *rational* thing to do. Hence I will, following Davidson, call these kinds of explanations *rationalisations.*[6] Not all explanations of intentional action are rationalisations. Explanations of an action as being done out of habit, for example, or as a result of ignorance or mistake, do not reveal the agent's own perspective on their action in this way. An explanation in terms of the agent's character, upbringing or culture might be more or less closely connected to a rationalisation. On the one hand, if the force of the explanation is to tell us something about what the person considers important, right or appropriate ('She took her shoes off because she grew up in Japan'), it might connect closely to the person's own perspective on what they are doing; on the other, if the point is simply that people of the relevant kind tend to behave in this kind of way ('He did it because he's a jerk'), it might not be.

In distinguishing rationalisations from other kinds of explanation, we might note another connection with Anscombe's discussion of intention, in which she marks out explanations of action that give reasons for that action as a distinctive class. At the beginning of *Intention*, this shows up in a discussion of expressions of intention as predictions of future behaviour. Sometimes, we predict our own future actions by relying on empirical generalisations about what we tend to do in certain circumstances. For example, I might predict that I will make a fool of myself at the party on the basis that I get drunk at parties and when I get drunk I usually make a fool of myself. If I tell you that I am going to make a fool of myself at the party, I might be making a prediction of this kind. On the other hand, I might be doing something different: I might be expressing my intention to make a fool of myself. Perhaps I have some reason for making a fool of myself, such as that it will endear me

[5]. The question can have application but only a negative answer, where the agent is doing what they are doing for no reason or no purpose. These are important problem cases for an account of intentional action, but not of primary interest to an investigation of rationalising explanation.

[6]. This is a technical use of 'rationalisation', importantly different from its more common everyday sense. In the present sense, a rationalisation is a genuine explanation of an action that shows what point the agent saw in taking it. In the everyday sense, a rationalisation is not typically a genuine explanation at all: it may give considerations that would have been good reasons for taking the course of action in question, but these are not also reasons *why* the action was taken. They are, rather, would-be reasons why—reasons the person might have acted on had they done the same action for better reasons than those which actually motivated them.

to the other people in the department. I have decided that I will make a fool of myself at the party, and when I say 'I'm going to make a fool of myself at the party', I am expressing my intention to do so. My statement, Anscombe says, is a genuine prediction, since it is a statement which 'later ... with a changed inflection of the verb, can be called true (or false) in face of what has happened later' (Anscombe, 1963, p. 2). However, it is importantly different from a prediction based on empirical evidence. In this latter case I do not arrive at my prediction applying anything like a theory of my own behaviour; rather, I think about what *to* do and the reasons in favour of taking the various options that are open to me. My prediction is based on practical, rather than theoretical, reasoning. Even though psychological prediction and explanation are not, I think, as symmetrical as is sometimes supposed,[7] a closely related distinction applies to explanations of a person's actions. Indeed, the very grounds upon which someone forms an intention may, after that intention has been carried out, explain their action in the way we are interested in.

So not just any explanation of an intentional action constitutes a rationalisation. Nonetheless, even genuine rationalisations form a diverse category. The most central cases seem to be those in which the person's action is explained in terms of either their desires, their goals, their beliefs, or by facts about the course of action itself or what might be achieved by it. These kinds of explanation are by no means exclusive, and explanations of each kind can together be true of a single action. Suppose, for example, that I am going to the Hereford market. The following explanations could all be true and would all rationalise my action—that is, show the point I see in what I am doing, and explain my action as (in some sense) rational:

1) They have Jersey cows there.
2) I want to buy a Jersey cow.
3) I think a Jersey cow would suit my needs.
4) I am going in order to buy a Jersey cow.

It is natural to say that a rationalisation gives us the agent's reasons for acting, or the reasons for which the agent did what they did, but this natural thought raises a puzzle. On one hand, the explanations above each give a very different kind of consideration: respectively, a fact about the market, a fact about what I want, a fact about my beliefs and a fact about my further intentions in acting. One fact about the world and three about my mind. On the other hand, there is some pressure towards thinking of reasons for acting as being all of a kind. Reasons for taking some course of action or another—'normative' reasons —are supposed not only to explain people's actions when people act for such reasons; they are also supposed to determine what an agent should or ought to do. They need to be the kinds of things that can be weighed against one another, that can stand in logical relations, and so on. The thought of weighing a worldly fact against a goal, for example, seems rather obscure. Unless we can give some account of how the considerations in (1)–(4) can all be reasons together, there seems to be a problem understanding how there can be any unity

[7]. See (Andrews, 2003).

between the different kinds of rationalisation. Notably, it seems plausible that some but not all rationalisations explain by giving the agent's reasons for doing what they did.

However, we should certainly try to maintain the idea that at least *some* rationalisations give the agent's reasons for acting. If people sometimes act on the basis of good reasons —'normative' or justifying reasons—presumably this means that they sometimes do what they do *because of* those reasons. Hence we should expect that some rationalisations do simply state the (normative) reason for which the agent acted. Practical reason, if it does anything, ought to make our actions at least somewhat sensitive to the reasons for them, so that we at least sometimes do the things we do because there is good reason to. Actions taken for reasons should be explained by those reasons.[8]

If we keep in view the idea that normative reasons sometimes motivate and hence explain actions, it will be clear that not much is to be gained by accommodating the diversity of rationalisations by distinguishing two different kinds of reason, normative and motivating, or perhaps normative and explanatory,[9] and saying that all rationalisations give reasons of the latter kind. Suppose first that we make a distinction between normative and explanatory reasons. The problem here is that two quite different readings of 'explanatory reason' suggest themselves, neither of which is particularly helpful. On the first reading, an explanatory reason is just any explanation. Explanations give reasons *why* things happen or why things are as they are, and we could perfectly sensibly calls these 'explanatory reasons'. In this sense, though, any explanation of an action, rationalising or not, gives an explanatory reason, so this conception of explanatory reasons will tell us nothing interesting about the the special character of rationalisation. On the other reading, we narrow the 'reason' in 'explanatory reason' in some other way. The most obvious way to do that is to say that explanatory reasons are just those normative reasons that also explain actions. If we understand it in this sense, though, all the questions with which we started still remain to be answered.

Now suppose instead that we appeal instead to a distinction between normative and motivating reasons. Here the situation is very similar to the situation with the second reading of 'explanatory reasons'. If we can act for good reasons, then we can be motivated by normative reasons; our motivating reasons can be normative reasons. Not all reasons that count in favour of a person's doing something actually motivate them to do it, of course, so it is not that the distinction between normative and motivating reasons is not real. The point is rather that if we can be motivated by normative reasons, then the distinction is one of role rather than kind:[10] a motivating reason is just a reason—a normative reason—that plays a motivating role. We could, of course, choose to say that any rationalisation gives a motivating reason, but this is apt to obscure what we are seeking to elucidate if the first-person perspective revealed by a rationalisation is a perspective on normative reasons and some but not all rationalisations give motivating reasons that are also normative reasons. It seems to me that a better approach will be to use 'motivating reason' to refer only to genuine normative reasons by which a person is motivated, and ask: Which rationalisations explain

[8]. (Heuer, 2004). Compare (Dancy, 2000).
[9]. (Smith, 1994) is one author who employs this tactic.
[10]. This point is made forcefully by both (Alvarez, 2010; Dancy, 2000).

by giving the agent's motivating reasons, and how do these rationalisations relate to each of the other kinds? In particular, how do the non-reason-giving rationalisations explain actions as rational if not by giving reasons for which the agent acted?

I will, in this thesis, attempt to lay some of the neccesary groundwork for understanding the unity and diversity we find here. Because each of our forms of rationalisation relates to each of the others in different ways, the task must be addressed piecemeal. I will begin by considering the relation between two forms of explanation that are particularly intimately connected, namely those that give 'worldly' facts about the action or the agent's situation, and those that give facts about the agent's beliefs. My reason for considering these first is that there is a good case for thinking that reasons for action are worldly facts, and that ascriptions of belief rationalise action in virtue of some kind of connection to such reasons. So explanations in terms of worldly facts are especially important because they are the explanations in which we explain someone's actions simply by stating their reasons for acting, and explanations in terms of the agent's beliefs are especially closely connected with these. In this chapter and the next I will attempt to make sense of the exact nature of this connection.

1.2 Some preparatory ground-clearing

Before moving on, I want to make some general comments about the ideas of fact, reason and rationalisation, making clear some of the assumptions upon which the subsequent discussion will (and will not) depend. Speaking in somewhat impressionistic terms, we can contrast two main approaches to the notion of 'fact'. On the first, facts are concrete entities, 'truth-makers', the things in virtue of which true propositions or statements are true.[11] On the second, facts are, broadly speaking, representational, perhaps just identical with true propositions, which might themselves be understood as 'logical constructions', for instance, or as whatever a true statement states.[12] I shall try for the most part to remain neutral about the nature of facts, though not because I think it irrelevant to the issues at hand. Questions about the nature of facts themselves potentially bear on a number of issues that will come up in the course of the investigation, such as the nature of explanation, of the way an agent relates psychologically to worldly facts, and to the relation between knowledge and belief.[13] Nonetheless, I will attempt to address such issues, where they arise, in a way that does not depend on a theory of truth or of facts.

If we understand 'fact' in this non-committal way, the idea that reasons for action are facts enjoys quite widespread support, even among authors with otherwise quite different views about the nature of normativity and practical reason.[14] While different authors

[11]. See for example (Austin, 1950).

[12]. See (Prior, 1971) and (Strawson, 1950) respectively.

[13]. For an example of the last of these, see (Hyman, 2017), who argues that a fact is not a true proposition but the truth of a proposition, hence that knowledge and belief do not have the same kind of content.

[14]. Noteworthy examples include, but are by no means limited to, (Alvarez, 2010; Collins, 1997; Dancy, 2000; A. H. Goldman, 2009; Hyman, 1999; Raz, 1986, 2000; Scanlon, 1998; M. Schroeder, 2007; Stampe, 1987; Williamson, 2000).

motivate this view in somewhat different ways, the basic idea behind it is quite simple. 'Normative' reasons for action are those considerations that favour or justify a person's taking one course of action or another (or refraining from some action). They bear on what a person ought to do in a given situation. Regardless whether we think what person ought to do is a matter of maximising happiness, doing what will satisfy their own desires, doing their duty, doing what is just, simply doing good, or something else besides, whether some course of action will meet the relevant standard of justification is an objective matter: it depends on the actual nature of the situation, what outcomes are actually possible or likely from the person's acting in the relevant way, and so on.[15] If we are interested in what a person should do, whether that person is we ourselves or someone else, the things to consider are the facts that bear on the matter. These are reasons for acting. If this is right, then to act on the basis of a reason for so acting is to act on the basis of a worldly fact, and the class of rationalisations that give genuine reasons for which the agent acted are what I will henceforth call *worldly rationalisations*: explanations that explain a person's action simply by stating some worldly fact that constituted a reason for them so to act, such as our (1) above.

The cogency of worldly rationalisation appears to depend on the universal form of judgements about worldly reasons. In a case of worldly rationalisation, a person's action is explained by a worldly fact itself, not by any idiosyncratic or particular feature or characteristic of the agent, except insofar as the latter must be recognised as conditions or circumstances relevant to the fact's itself constituting a reason. The idea that worldly reasons can themselves explain actions goes hand in hand with the idea that when the fact that *p* is a reason for a given person to *A*, the fact that *p* would be a reason to *A* for anyone in relevantly similar circumstances. Another person recognises the fact that *p* as a reason to *A* and *A*s as a result; we recognise the fact that *p* as a reason to *A* and so understand this person's *A*-ing.

The universality of worldly reasons and rationalisations gives us a new way of looking at the questions raised at the end of the previous section. Worldly reasons are universal, and worldly rationalisations seem to exploit this universality. The various other forms of rationalisation we considered above, though, explain actions by citing not universal reasons but idiosyncratic features of the agent themselves. How, if at all, do such rationalisations relate to worldly reasons and rationalisations? In particular, does making sense of them require us to amend our conception of reasons as universal? I will argue that rationalisations in terms of the agent's beliefs—what I will call, for reasons that will become clear, *perspectival* rationalisations—can be seen not to present such a challenge: their rationalising role can be quite straightforwardly reconciled with the universality of reasons. However, as we will see later, things are not so simple when we think about certain other kinds of idiosyncratic rationalisation.

Finally, a word about the scope of rationalisation. So far the discussion has been solely focused on the rationalisation of action, and this will be the primary focus throughout most of this thesis. We should acknowledge, however, that actions are not the only things to

[15]. For the compatibility of the idea that reasons are facts with the claim that they have subjective conditions, see for example (A. H. Goldman, 2009; M. Schroeder, 2007).

which this special kind of explanation—explaining-as-rational, explaining something by showing how it made sense from the agent's or subject's point of view—applies. Intentions and decisions, of course, can be rationalised, presumably by the same kinds of considerations that rationalise actions. More significantly, beliefs, judgements, and arguably many emotions can also be rationalised, here by different kinds of considerations corresponding to the different kinds of reasons that favour each of the responses in question.

Some of this will be significant in what is to come. In the next chapter I will address the question of the relation between worldly rationalisation and perspectival rationalisation, and I will do so in part by considering the way in which beliefs of a certain kind are themselves rationalised. It is important, then, to note that the dual scheme of worldly and perspectival rationalisation applies to beliefs in much the same way as it does to actions. We typically expect someone's beliefs to be responsive to reasons, to be held for good reasons and to be revisable in the light of new evidence, and so on. As with action, we can sometimes explain someone's believing something by reference to something other than their reasons for believing it, such as their character, their upbringing, that they were taught it in school, that they haven't really thought about it, and various other sorts of facts about them, their psychology, and their history. Two central ways of explaining why a person believes what they believe, though, are worldly rationalisation and perspectival rationalisation—citing actual reasons on the basis of which the belief is held (He thinks the secret police are after him because they are cracking down on dissidents and several of his comrades have recently disappeared), and citing other beliefs on the basis of which the belief being explained is held (He thinks the secret police are after him because he believes he has uncovered evidence of a global reptilian conspiracy).

Similarly, a person's wanting something or feeling a certain way (sad, happy, angry, ...) can also often be rationalised. Sometimes it is not clear whether something of this sort can be rationalised. This is an issue to which we will return in Chapter 3. Nonetheless, there do seem to be relatively clear examples of rationalisation here, as when I say that I want a new pair of shoes because my shoes are all boring and plain, or that Karoline was distraught because she thought someone had stolen her phone.

These issues, then, are not restricted to the case of action. Nonetheless, the primary focus of this thesis will be on the rationalisation of action. One reason for keeping a narrow focus is that what makes some fact a reason for someone to believe something or to feel a certain way will be quite different from what makes a fact a reason for them to perform some action. Different normative standards apply in each case, and so considering all possible targets of rationalisation is liable to introduce a great deal of complexity. Another reason is that the existing literature on reasons and rationality, and on psychological explanation, focuses largely on the case of action. The explanation for this may be partly historical (the influence of behaviourism, perhaps), but I suspect there are also some quite basic non-historical reasons for it, such as the central importance of understanding others' actions to human life and cooperation, and the more essentially 'public' nature of actions as events that we can witness first-hand.

The most basic reason for my focus on action, though, is that the central claim of the thesis, concerning the universality and particularity of reasons, is most clearly made by looking at the case of desire, the relation of desire to worldly reasons, and the way in which we understand our own actions in relation to certain of our desires. This, plainly, is specific to the case of action. Whether the point can be generalised in any way to the rationalisation of beliefs, emotions and so on is a nice question, but not one that I will attempt to address.

1.3 Perspectival rationalisation: the puzzle

We have identified two different kinds of rationalisations. A worldly rationalisation makes sense of an action from the agent's point of view by citing a fact about the world that constituted a reason for the agent to act as they did and which was the agent's reason for so acting. A perspectival rationalisation explains an action by citing a belief of the agent. I will return to the issue of worldly rationalisation, and its relation to perspectival rationalisation, in the next chapter. For now, I want to consider a different question, namely the relation between perspectival rationalisation and *reasons*, which we have provisionally identified with worldly facts.

Rationalisations make sense of actions as rational. They explain actions by showing us what rationally motivated the agent. If reasons for acting are facts, then it seems that what rationally motivates an agent is not always a reason. When someone acts on the basis of a false belief, the consideration that rationally motivates them, the consideration which, from their perspective, gives their action a point, is false, and hence not a fact. In such cases, we cannot correctly give a worldly rationalisation of the person's action. If there are no Jersey cows at the Hereford market, it cannot be true that someone is going there because there are Jersey cows there. We have to retreat, as it were, to the perspectival rationalisation: She is going there because she thinks there are Jersey cows there.

If we remain strict in our use of 'reason', we may have to say in cases such as this that the agent acts for no reason. As Simon Blackburn says, this might 'sound harsh', given that the agent was not irrational and 'certainly had their reasons for what they did, and … may have acted well in the light of them' (Blackburn, 2010, p. 8). Indeed, this has even been raised as a challenge to the identification of reasons with facts. Juan Comesaña and Matthew McGrath, for instance, claim that whenever an agent acts rationally they act for a reason, and take this claim to be obvious enough that they feel entitled to use it as an unargued-for premise in an argument for the view that some reasons are false and hence that not all reasons are facts (Comesaña & McGrath, 2014).

Such concerns might be taken to motivate some claim to the effect that the agent's reason is, in such cases, her belief. As Alvarez (2010) has stressed in this context, 'belief' is ambiguous between (i) what a person believes (a proposition or belief content) and (ii) that person's believing it (a state, attitude, or psychological fact). Hence the idea that the agent's reason is her belief is ambiguous between the claim that her reason is what she believes—

which is, in the relevant cases, a falsehood—and the claim that her reason is the state or fact of her believing.[16] The two possibilities are each, in different ways, unsatisfactory.

Let us first consider the second suggestion, that where we give a perspectival rationalisation of the action the fact that the agent has the relevant belief is their reason. The issue here is that while the fact that I believe something can be a reason for me to act, it is not a reason for the kinds of action that it rationalises when the form of rationalisation we give is perspectival. This point is best brought out by way of illustration. Suppose that I believe that the secret police are after me.[17] My believing this—the fact that I believe this—might be a (worldly) reason for me to do something, for instance to see a psychiatrist, and I might do that very thing for this very reason. Here, the fact that I have this belief is my reason for going to see a psychiatrist. This is just another case of worldly rationalisation; it just happens to be one in which the worldly reason for my action concerns the state of my mind. We have a perspectival rationalisation only when the relation between belief and action is of a different kind: as when, for instance, I flee the country because I believe that the secret police are after me. If the secret police are *not* after me, then although my believing this explains my action in the rationalising kind of way, there is a truth that is clearly expressed by saying that there was not in fact any reason for me to flee the country and that I therefore fled the country for no reason. The latter locution, 'He did it for no reason', is apt to be misleading because we often use it to mean that someone acted for no purpose or for no point. In that sense, it would not be correct to say that I act for no reason when I flee the country on the basis of the false belief that the secret police are after me: there is clearly a point to what I am doing, at least from my point of view, that point being to avoid capture and persecution. This point is revealed by the perspectival rationalisation.[18] If we are just a little stricter with our use of 'reason', though, we can say that while I acted with a (somewhat) rational purpose, still I did not act for a (genuine) reason.

There is perhaps more to be said for the other reading of the claim that the agent's reason is her belief. There is a sense in which when I act on a false belief, my belief (the thing I mistakenly believe) plays a reason-like role in my rational psychology. It figures in my perspective on my action in much the same way that a worldly reason would if I were acting for a worldly reason. It gives the rational basis for my action; it is that in the light of which I act.

One way to think of this is that what I believe functions as a premise of my practical reasoning, which for our example we might represent something like this:

- The secret police are after me.
- If the secret police are after me, I can avoid capture by fleeing the country.

So I flee the country.

Each of the premises is something I believe and which may or may not be true. My ability to engage in this kind of practical reasoning and so to act rationally on the basis of

[16]. Strictly speaking (ii) contains two alternative claims: that the reason is the belief-state, and that it is the fact of the agent's believing. See (Dancy, 2000, Chapter 5) for discussion of this distinction.

[17]. I take the example and its use from (Hyman, 1999); see also (Dancy, 2000, Chapter 6).

[18]. See (Alvarez, 2010, pp. 141–7).

such premises does not depend upon those premises being true. So how can the truth or falsity of my beliefs make a difference to whether I am acting for a reason?

While we should take this line of reasoning seriously, there is good reason not to conclude that in false belief cases the agent acts for a reason. First, it is not a normative reason: it cannot justify the action. A falsehood is not the kind of thing we should consider in figuring out what to do. Second, although it plays an important role in making the action rational, there is a sense in which it cannot itself make the action rational. This is connected with the fact that it cannot rationalise the action—it cannot explain it as rational, because it cannot explain the action at all. Whenever we have an explanation, a canonical explanatory statement should be possible—a statement of the form '*p* because *q*'. If this canonical statement is to be true, then it must be true both that *p* and that *q*. Hence both what is explained and what explains it must be facts. When someone acts on a worldly reason which is the fact that *q*, simply stating their rational grounds for acting can satisfy this requirement: 'He *V*-ed because *q*'. When someone acts on the false belief that *q*, though, the requirement is not satisfied: 'He *V*-ed because *q*' will be false. What will satisfy the requirement is a perspectival rationalisation: 'He *V*-ed because he thought that *q*'. It seems that while a false idea can motivate, it cannot explain.[19]

Of course, if I flee the country because I think that the secret police are after me, it may seem to me that I am fleeing the country because the secret police are after me. This only shows that when I am mistaken about how things stand I may, insofar as I act on my mistaken belief, also be mistaken in a certain way about why I am doing what I am doing. In another way I will not be mistaken, because I will know that I am fleeing the country because I think that the secret police are after me. In this respect I am not mistaken about the rational grounds of my action; I am not mistaken about my practical reasoning. But if I think that I am fleeing because they really *are* after me, this is a mistake. This is something we all implicitly acknowledge. If I discover my error and realise that I was misguided, I will not continue to insist that I fled because they were after me. I will, correctly, advert to the fact that I believed that they were.

To say that beliefs are reasons, or that we can have false reasons, would muddy the crucial distinction between those cases in which someone's doing something can be rationalised by there having been a reason for them to do what they did, and those cases in which we need, in order to rationalise the action, to say something about what the person thought was the case. When the agent is mistaken, it is not *what* they believe that explains their action as rational. If someone asks what makes my fleeing the country rational, 'The secret police are after him' cannot be a correct answer if the secret police are not in fact after

[19]. (Dancy, 2000, Chapter 6) argues on the basis of false-belief cases that not all explanation is factive in the way I am claiming. In particular, Dancy claims that where the considerations in light of which a person acts are false or do not obtain in reality, those considerations nonetheless explain the person's action. I think Dancy is correct about something here, namely that even when someone acts on a false belief, there is a sense in which we still *understand* their action in terms of the considerations on which they acted—considerations which are, in such cases, false. However, I think we can acknowledge this without endorsing the claim that falsehoods can explain. This will hopefully become clearer in the next section, when I present my positive account of how perspectival rationalisations rationalise.

me. What makes my action rational is something to do with the role this proposition plays in my practical thought. This is why the correct, perspectival rationalisation is given by a fact about my psychology. Nonetheless, I believe that the right account of how perspectival rationalisations explain will have to accommodate what truth there is in the idea that when someone acts on a false belief, it is what they believe that makes sense of their action from their point of view. The key, I will suggest, is in the last part of this idea: the idea of the agent's point of view.

1.4 On the agent's point of view

In *The Quest for Reality*, Barry Stroud, discussing the fact that the contents of psychological attitudes are 'typically specified in terms which mention only circumstances that do or could hold in the nonpsychological world', says:

> [W]e who inhabit the world can understand someone in that world as believing something or as perceiving something only if we can somehow connect the possession of the psychological states we attribute to the person with facts and events in the surrounding world that we take the beliefs and perceptions to be about. We understand one another to be parts of, and engaged in, a common world we all share. If we ourselves had no beliefs at all about what is happening in the environment or what another person is likely to be paying attention to, we would be in no position to attribute any beliefs or perceptions to that person at all. So it looks as if we interpreters and ascribers of beliefs and other psychological states must be engaged in the world, in the sense of taking certain nonpsychological things to be true of it, if we are ever going to attribute beliefs or perceptions to anyone.
>
> In identifying the contents of the attitudes we ascribe, we must inevitably start with what we already know or believe, or can find out, so we have no choice but to attribute to others, at least in general, beliefs in and perceptions of the very things we ourselves take to be true or to exist in the world. We cannot make sense of someone as believing something we know to be false unless we can identify what he believes and can offer some explanation of how he comes to get it wrong. That involves attributing to the person many other beliefs, the possession of which helps make his particular divergence intelligible. And those further beliefs will typically include many that we share. Those we do not share will, in turn, be attributed only if we can understand how a person inhabiting and reacting to the world we all live in nonetheless came to have them. (Stroud, 2000, pp. 150–1)

When we rationalise people's actions, we are seeking to understand how their behaviour makes sense in the circumstances of the world we all share. I have suggested that we understand reasons for action as facts about the world as it really is. We seek to act in ways that are justified by such reasons. These reasons are universal; they do not depend on the agent's idiosyncratic perspective. When someone acts on the basis of a false belief, though, we need to recognise a form of idiosyncrasy in order to make sense of their action.

Making sense of someone who acts on the basis of a false belief involves seeing things from the agent's point of view whilst recognising that that point of view is mistaken. But how can a mistaken point of view explain an event in the real world? This puzzle, or something close to it, is nicely articulated, albeit in passing, by Bernard Williams, in 'Internal

and External Reasons' (Williams, 1981a). Williams famously describes a case in which a person wants a gin and tonic and believes the stuff in the bottle before him is gin when in fact it is petrol. Williams expresses our present difficulty as follows:

> On the one hand, it is just very odd to say that he has a reason to drink this stuff, and natural to say that he has no reason to drink it, although he thinks that he has. On the other hand, if he does drink it, we not only have an explanation of his doing so (a reason why he did it), but we have such an explanation which is of the reason-for-action form. (Williams, 1981a, p. 102)

It seems pertinent here to distinguish two possible uses of 'rational', one in which the agent would be rational in drinking the stuff before him and one in which he would not. In an objective sense, it would be irrational for the agent to drink the stuff before him; there is no reason for him to do it and a very good reason not to, namely that it will make him ill. In the subjective sense, on the other hand, it would be quite rational for the agent to drink the stuff, since he wants a gin and tonic and believes the stuff to be gin. From his point of view, it is the most reasonable course of action.[20] There are corresponding objective and subjective senses of 'ought'. The agent in Williams's example objectively ought (if he wants to stay in good health) not to drink the stuff, and he subjectively ought (if he wants a gin and tonic) to drink it. When an agent acts for a good reason—when they act in such a way that their so acting can be explained with a worldly rationalisation—objective and subjective rationality coincide. This is not the case with perspectival rationalisations: a perspectival rationalisation makes the action intelligible merely as subjectively rational. To understand how perspectival rationalisations work, we need to say what this means.

1.4.1 Subjective rationality

Niko Kolodny (2005) offers a way of thinking about subjective rationality which seems particularly amenable to the idea that truth is privileged in our understanding of others. Kolodny argues that we should see the normativity of subjective rationality as, in a sense, merely apparent. He presents a 'transparency account' of subjective rational ought-statements, on which statements about what an agent subjectively ought to do are in effect statements about the agent's perspective, rather than statements about what the agent actually ought to do. On this view, when we say that Bernard, given that he believes the stuff is gin, ought to drink it, what we mean is that, as things seem to Bernard, he ought to drink the stuff. Saying this is consistent with insisting that, as things actually are, he really shouldn't drink it. The mismatch between the 'ought'-statements is no more than the mismatch between the agent's point of view on the world and the way the world really is. Understanding an action as rational is just understanding that action as something that seemed, from the agent's perspective, the thing to do. The real question will then be how we

[20]. (Kolodny, 2005) uses the terminology of objective and subjective rationality in making the same distinction. See also the distinction between substantive and structural normative claims in (Scanlon, 1998).

should think about 'the agent's perspective', and what is involved in *taking* the agent's perspective, which coming to understand their action in this way would presumably require.

First, though, Kolodny's account requires some extension and refinement if it is to form the basis of a satisfactory account of psychologised rationalisation. There are two significant limitations to Kolodny's account as it stands. First, his focus is exclusively on rational *requirements*, and so he only gives an account of the subjective rational 'ought', which he understands in terms of its seeming to the agent that they have conclusive reason to do or not to do something. This may be fine for Kolodny's aim of explaining the apparent normativity of rationality and the idea of a subjective rational 'ought'. Our concern, though, is with rationalising explanation, and for this we need a conception of rationality that is somewhat weaker. We do not act only in ways that we are rationally required to, and in general a rationalisation does not explain an action as something that the agent rationally *had* to do. A rationalisation merely shows what point the agent saw in acting as they did, and to be a rationalisation it need only rationalise in this thin sense. For this reason, having a correct rationalisation is often consistent with there being further questions about whether the action rationalised was, in a more demanding sense, the rational thing to do. Moreover, even a rationalisation that does show an action to be rational in a more demanding sense need not necessarily do so by showing that the agent was rationally required to take that action. Very often we find ourselves with a range of 'eligible' options, for each of which we have sufficient but not conclusive reason, such that choosing any would be objectively rational.[21] If we recognise our situation as such or if it seems that way to us, then we may in the same way be perfectly subjectively rational in choosing any of the relevant options.

Kolodny's account is of subjective rational requirements, and his claim is that when we say that an agent ought rationally to *V*, we mean that, as it seems to the agent, they ought rationally to *V*; they believe they have conclusive reason to *V*. So if we have a rationalisation of the form 'A *V*-ed because she thought she had conclusive reason to *V*', Kolodny has an account of the sense in which this explains the action as rational. The account can easily be extended simply by saying that a perspectival rationalisation shows us that the agent believed that they had some reason to do what they did. In the cases where the rationalisation shows the action to be rational in the more demanding sense, it shows that the agent believed that they had sufficient reason to act as they did.

This brings us to the second unsatisfactory aspect of Kolodny's account, which is that it is too intellectualist for our purposes. On Kolodny's picture, the (subjective) rationality of an action for an agent is determined by the agent's beliefs about their reasons as such. The explanations that we are interested in do not characteristically attribute beliefs about reasons as such; they typically attribute ordinary beliefs about how things are, for example the belief that they have Jerseys at the Hereford market. The belief that explains my going to the market is not the belief that I have a reason to go to the market, but a belief whose content might, were it true, constitute or correspond to a reason for me to go to the market.

[21]. This is a major theme in the work of Joseph Raz. See in particular (Raz, 1986, 2000). For a range of perspectives on this issue, see the articles collected in (Chang, 1997a).

It might be suggested that a perspectival rationalisation like 'He went to the Hereford market because he thought that they had Jersey cows there' explains by indicating something that the agent believed to be a reason, and that we come to appreciate the agent's perspective on his action when we infer that he also believed that this was a reason. This seems to me to be misguided. It is not so much that I think we do not have beliefs about our reasons as such. Perhaps we do, and perhaps this is even an essential part of acting for a reason. Perhaps not.[22] The issue is that what we are seeking to understand is our understanding of actions from the agent's point of view. This involves coming to appreciate the specific point that the agent saw in doing what they did. If I find you laying out all the green objects in your house on your roof, I might find it interesting to learn that you believe there is a good reason for you to do so, but this will not help me to understand you: for that I need to know what the putative reason actually is.[23]

Moreover, believing that some consideration is a reason to act in a certain way is not enough for your belief in that consideration to make your action intelligible. There are limits on what can intelligibly be taken as a reason for what. To borrow an example from Raz, I cannot (intelligibly) choose to have coffee because I love Sophocles (Raz, 2000, p. 8) and that is the case even if I believe that my love for Sophocles is a reason to have coffee. If some further factual beliefs were added that made sense of my love for Sophocles' being such a reason, my choice might become intelligible. Perhaps I am under the impression that Sophocles loved coffee, and that by drinking coffee I will be honouring Sophocles. You might wonder how I acquired this odd notion, of course, but so long as I have it, it seems enough to make sense of my having the coffee. What appears to be doing the work here, though, is not my beliefs about reasons for action as such, but rather the way in which what I believe would, if true, actually show some worth in doing what I am doing.[24]

Kolodny considers a suggestion along these lines, which he credits to Pamela Hieronymi and Seana Shiffrin, with respect to his position on rational requirements. The suggestion is that

> when we say 'You ought to *A*; it would be irrational of you not to,' we mean not (in general) 'As it seems to you, you ought to *A*,' but instead, 'As it seems to you, something is so, and (although you may not have realized it) if that is so, then you ought to *A*.' (Kolodny, 2005, n. 47)

Kolodny says he has one misgiving about this suggestion, which is that he doesn't think someone is irrational in failing to *A* merely because they are not aware that something they believe would be conclusive reason for them to *A* if it were true. The misgiving may be onto something. Suppose that some ordinary person believes some mathematical truths which entail, by way of a complex proof, the truth of Fermat's Last Theorem. Suppose this is

[22]. See (Lavin, 2011) for helpful discussion of this issue.

[23]. There is also a good case for thinking that what it is rational for us to do is not as straightforwardly determined by our beliefs about our reasons as such as Kolodny's account suggests. See for instance the discussion of 'inverse akrasia' in (Arpaly, 2002).

[24]. Compare Anscombe's claim that 'the good (perhaps falsely) conceived by the agent to characterise the thing [they want] must *really* be one of the forms of good' (Anscombe, 1963, pp. 76–7).

enough to make those mathematical truths conclusive reasons to believe the theorem. Being an ordinary person, our agent simply has no way of knowing that some things they believe would, if true, entail Fermat's Last Theorem. It would be absurd to accuse this person of being irrational for failing to believe the theorem; given that they cannot see the entailment, it would normally be irrational for them to believe it. Or, suppose that someone mistakenly, but through no fault of their own, believes that someone who is drowning will splash around and call for help. If this person sees a bather bobbing up and down, with their arms extended laterally, pressing down on the surface of the water, they might not realise that this is what the instinctive drowning reaction looks like and that the person's exhibiting this behaviour is therefore a conclusive reason to go to their aid. This might be tragic, but, given the agent's ignorance, it would be unfair to accuse them of irrationality. So there does seem to be a problem with the proposed revision. On the other hand, Kolodny's suggestion seems unpalatable as well, for the reasons given above.

There is a way to address the worry that Kolodny raises without committing to his intellectualist account, which is to appeal to the agent's *competences* to respond to reasons of relevant kinds, without taking a stand on whether such competences have to involve explicitly normative beliefs. An account of this kind is developed by Kurt Sylvan. Sylvan's account is expressed in terms of *apparent reasons*, which are meant to be the kinds of things to which a rational agent responds—essentially, the contents of the beliefs cited in perspectival rationalisations.[25] Without going into too much detail, the basic idea of Sylvan's account is that an agent has an apparent reason only when they have a 'relevant reasons-sensitive competence' to respond to reasons of the relevant kind. The idea of a relevant reasons-sensitive competence is fleshed out in terms of objective (that is, worldly) reasons: an agent has an apparent reason R to *V* (and hence a belief in R that is apt to rationalise their *V*-ing) only when they have 'a competence to treat R-like considerations like objective reasons to do [*V*]-like things only if they are, when true, objective reasons to do [*V*]-like things' (Sylvan, 2015, p. 599). This in effect allows us to remain neutral on the issue of intellectualism: perhaps treating an R-like consideration as an objective reason involves believing that R-like considerations are objective reasons, but perhaps it does not. Applied to our example above, an account along these lines would allow us to say that the non-mathematician is not irrational in failing to believe the truth of Fermat's Last Theorem because, although he has conclusive reason to believe Fermat's Last Theorem, he could not believe it *for* that reason, because he lacks the mathematical skill to run through the relevant proof. In general, we can say that a belief-ascription can rationalise an action when what the agent believes would, if true, constitute a reason for them to act as they did, and the agent had the competence to treat considerations of that kind as reasons to respond in the relevant way.

[25]. (Sylvan, 2015); on 'apparent reasons' see also (Alvarez, 2010).

1.4.2 Taking the agent's point of view

So we have a rough picture of when ascribing a belief, even a false belief, can show an action to be rational. The basic idea about subjective rationality that we have in hand is one on which ascribing a belief to an agent explains their action by showing that, from their perspective, it was something that made some kind of sense for them to do. Understanding an action as subjectively rational, then, involves, in some sense, coming to see things from the agent's perspective. I want to suggest that understanding another's action in this way involves employing our own competence to respond to reasons—to respond to worldly reasons—within a counterfactual or suppositional context. When we understand an action on the basis of a worldly rationalisation, we do so by recognising how, had things been as the agent took them to be, there would have been a reason for them to do what they did.[26] Although perspectival rationalisations explain actions in terms of an idiosyncratic feature of the agent, then, the character of our understanding is fundamentally based on our and the agent's shared grasp of universal, worldly reasons.[27]

If we are not competent, or are not disposed, to think about the same kinds of facts in the same way as the agent did, then this way of understanding the agent's doing what they did will not be immediately available to us. A standard perspectival rationalisation, which simply cites the belief on the basis of which the agent acted, will not be enough to reveal to us their perspective on their action. We might need some further explanation. Often this will consist not in further explanation of the agent's thinking as such, but simply explanation of how what the agent believed would if true have given them a reason to do what they did. This same structure can occur with respect to worldly rationalisations. For instance, the person in our example above who is ignorant about what drowning looks like might not understanding why a lifeguard dashes out to the bather's aid. Saying 'They were bobbing in the water with their arms out to the sides, pressing down on the water's surface' will not be an adequate rationalisation for this person: it will not enable them to see the point in what the lifeguard did. However, if we explain to them that this is what drowning looks like, they will understand, precisely because they thereby acquire the competence to respond to the relevant fact as a reason for the relevant kind of action.

More interesting cases are those in which a rationalisation fails to make someone intelligible to us because we do not understand the agent's values. Here, coming to understand the person might require not just further explanation but something more like training or acculturation. How exactly we should understand these cases leads into difficult questions about the metaphysics of values. Is 'learning' the values necessary to understand the other in such cases a matter of acquiring knowledge, or just of changing one's attitudes? Might there be cases in which the values necessary to understand the person are fundamentally inaccessible to us, so that they will necessarily remain unintelligible? These

[26]. Compare the notion of 'teleology in perspective' developed in (Perner, Priewasser, & Roessler, 2018; Roessler & Perner, 2013).

[27]. This kind of understanding thus consists in a kind of 'co-cognition' as characterised by (Heal, 2003).

questions are too large to address here.[28] Thankfully, we can develop a basic picture of rationalisation whilst remaining largely neutral on such issues. A broad range of views about the metaphysics of values can agree on the aspects of reasons that are central to the picture I am developing, such as that reasons are worldly facts and that reason-judgements are universal. Where they will differ is at a deeper level, for instance on what kind of thing our understanding of an action from the agent's point of view consists in, and on how to think of 'competences' to respond to reasons. Perhaps understanding others is a matter of our joint exercise of the faculty of reason. Perhaps it is a shared recognition of 'robustly real' values or normative principles. Perhaps it is no more than a certain sort of similarity or harmony in our responses to the facts in question. On any of these ways of thinking, we can see perspectival understanding as involving the exercise of shared capacities to respond to worldly reasons.

On the present account, grasping either a worldly or a perspectival rationalisation requires us to share the agent's perspective, in a certain sense. In the case of worldly rationalisation, you share the agent's perspective in that you both see things as they are—you are aware of the same reason and both recognise it as such. The problem presented by action based on a false belief is that if A *V*-s because she thinks that *p*, and you know that not-*p*, you cannot come to understand her action by thinking as she did about the reason there was for her to *V*, because there is no such reason. But you cannot come to understand the action by coming to *share* the agent's belief, if that belief is false: coming to think that A *V*-ed because *p* is not coming to understand A's *V*-ing if it is not the case that *p* and hence it is not the case that A *V*-ed because *p*. To echo Stroud, we want to understand the agent's behaviour in the circumstances they are actually in, in the common world that we share. However, we also do not want the kind of understanding provided by perspectival rationalisation to be utterly different in kind from that provided by worldly rationalisation. The solution I am suggesting is that we can take on the agent's perspective in a limited way, whilst keeping in view the facts as they really are, by supposing or imagining that things are as the agent mistakenly takes to be. Within the scope of this supposition, we can think about what there would be reason to do if things were that way. In doing so, we replicate or re-enact the agent's reasoning and thus come to appreciate the point they saw in acting as they did, without losing sight of the respect in which they were mistaken.

The idiosyncrasy introduced by perspectival rationalisation, then, is no challenge to the universality of reasons. A false belief is an idiosyncratic take on how things stand and hence on what reasons there are, but the understanding that a perspectival rationalisation can provide itself depends upon our and the agent's understanding of universal, worldly reasons. This account leaves open the relation between perspectival and worldly rationalisations. There seems to be an asymmetric dependence between the two forms: it appears that whenever a worldly rationalisation ('A *V*-ed because *p*') is true, a corresponding perspectival rationalisation ('A *V*-ed because she thought that *p*') is also true, whereas a perspectival

[28]. For a range of views that seem to me more or less consistent with the account being developed here, see for example (Blackburn, 1998, 2010; Enoch, 2011; Gibbard, 1990; Korsgaard, 1996; Raz, 2000, 2003; Williams, 1985).

rationalisation can be true without there being any corresponding worldly rationalisation. This raises a question about whether worldly rationalisation really *adds* anything to our understanding over and above what is provided by the corresponding perspectival rationalisation: is there anything special about someone's action being explained by a reason itself? In the next chapter we will consider some arguments on either side of this issue. First, in the remainder of the present chapter, I want to address a concern that might have arisen from the discussion up to this point.

Appendix: Is psychological explanation constitutively normative?

In this first chapter, I have attempted to articulate the beginnings of an account of a certain kind of explanation of action—what I am calling 'rationalisation'. I have emphasised two features of rationalisation in particular: that the understanding of an action provided by a rationalisation characteristically reflects the agent's self-aware understanding of their own action, and that rationalisations explain actions *as rational*.

In insisting that these two features of rationalisation characterise a certain kind of explanation of human behaviour, the present account is in tension with a way of thinking about the way in which mental states explain actions that might seem to be entailed by a very popular way of thinking about the nature of mental states and mental state-concepts. On the view of explanation in question, we understand the behaviour of human beings in just the same kind of way that we understand the behaviour of any kind of physical object. We have a causal theory consisting in generalisations about what kind of behaviour tends to result given certain causal conditions, and we understand the behaviour of physical objects by subsuming their behaviour under such a theory. When our subject-matter is human beings, the relevant theory consists largely of generalisations about how different kinds of mental states interact with causal inputs to generate behavioural outputs. We understand human behaviour, fundamentally, as what results from the causal interaction of input from the senses with 'internal' mental states like belief and desire. This is a picture of psychological explanation in which the first-personal and the rational do not appear to play any essential part.

One reason for favouring a view of this kind is that it might seem to be implied (or, perhaps better, presumed) by a very popular way of thinking about the nature of mental states or mental state concepts. *Functionalist* theories hold that mental state terms can be defined implicitly by the functional role that they play in a theory of a psychological system as a whole.[29] Such theories, it is generally supposed, will define that role in just the kind of causal-theoretical terms mentioned above: a given mental state's functional role is a matter of how it causally interacts with other mental states and with sensory 'inputs' and behavioural 'outputs'. Such functional generalisations might say, for example, that someone who desires that p and believes that if they V then it will be the case that p will tend to V, that someone who is in a position to see that q will tend to believe that q, and so on. On a standard functionalist approach, we derive implicit definitions of mental state terms by concatenating

[29]. Specific functionalist theories differ over whether the 'psychological system' in question is 'the mind', construed in the abstract, or the mind of a particular individual.

all the relevant generalisations (which together constitute our theory of the mind) and constructing a so-called Ramsey sentence by replacing each distinct mental state term with a different predicate variable and binding those variables with existential quantifiers.[30] What it is for something to be a belief, on this view, is just to play the causal role that belief plays in the theory. Whatever plays this role, functionalists tend either to argue or to assume, will be a physical state of some kind.

There are many different varieties of functionalism. One dimension of variation is in their interpretation of the ontological import of the functional theory—for instance, whether they take mental states to be identical with whatever 'first-order' state or property plays the role defined by the psychological theory, or instead with the 'second-order' state of *being in a state* that plays the relevant role. A dimension of variation that is more important for present purposes is in what the functionalist takes to be the theory of the mind that is relevant for definition mental state terms. So-called psychofunctionalist[31] theories hold that the theories that define the roles of mental states are the empirical psychological theories devised by cognitive scientists, whereas 'analytic' or '*a priori*' functionalists argue that the relevant theory is 'folk' or 'commonsense' psychology—represented, in Lewis's version, by 'all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses' (Lewis, 1972, p. 256).

It is not clear to what extent psychofunctionalism is relevant to the present discussion at all,[32] because a psychofunctionalist account could, it seems, be entirely silent regarding ordinary psychological explanation. If the attitude of belief has a causal nature that can be discovered by cognitive scientists and implicitly defined by theories in advanced cognitive science, this would not necessarily speak against the account we have been developing of how facts about what a person believes figure in our ordinary non-scientific understanding of that person's actions.

Analytic functionalism, on the other hand, might seem to be tied to the kind of picture of ordinary psychological explanation sketched above. If the idea is that the theory that implicitly defines 'belief' is a merely causal 'folk theory' of the mind, this would seem to be at odds with the view that rationalisation is a special kind of explanation that makes sense of actions by enabling us to see how they made sense from the agent's point of view. On the 'merely causal' picture, there is nothing particularly special about psychological explanations; they are causal explanations that make sense of certain events by subsuming them under causal generalisations, and in this they are just ordinary causal explanations of events of a certain kind.

A variety of options are available to avoid this worry. Two in particular are worth noting. First, we might argue that something about the character of rationalisation reveals functionalism to be false. Second, we might argue that while analytic functionalism might give the right account of how our mental state concepts should be defined, any adequate

[30]. See (Lewis, 1972) for an influential development of this approach.

[31]. This term originates in (Block, 1980); for an example of such a theory see (Fodor, 1968).

[32]. Except insofar as it might lead to the view that there is no such thing as belief. See (Stich, 1985).

functionalist theory will have to make room for the features that our discussion of rationalisation has highlighted. I will give a brief sketch of how each of these options might be pursued.

Davidson and the constitutive ideal of rationality

Donald Davidson famously argued that explanations of action in terms of mental states like belief and desire are essentially, constitutively, rational, and that as a result such explanations could not be understood in terms of their subsuming the action under a causal law. For Davidson, a rationalisation is a causal explanation in that it picks out something that was the cause of the agent's action—this is the difference between (a) having a belief that would rationalise an action and performing that action, and (b) performing that action because one has that belief (Davidson, 1980b, Chapter 1)—but the generality under which the explanation explicitly subsumes the action is not itself a causal law. On Davidson's view there is, for any two events related as cause and effect, a strict, exceptionless causal law under which they fall, but only as described in *physical*, not psychological, terms.[33] Psychological concepts are, for Davidson, irreducibly rational and normative.

Davidson's idea of radical interpretation plays an important role here. Very roughly, its significance is that things like rationality and truth play a constitutive regulating role in our ascriptions of psychological attitudes to others.[34] An agent's beliefs must be for the most part true, and their beliefs and actions for the most part rational, if they are to be intelligible to us as rational agents at all. Psychological concepts, though, are also causal, and rationalisations are causal explanations in that they identify causal conditions of the action that is explained. Because psychological interpretation is also constitutively rational, though, application of psychological concepts must be sensitive to an extra set of standards—standards of rational interpretation—that merely causal theories are not constrained by. One way this issue can be illustrated is through the problem of so-called deviant causal chains.

Someone might want to achieve some end and have a belief about how to achieve it, and might do the thing they believe would achieve the end, but not for that reason. This is why, Davidson argues, it is a necessary condition for someone's acting on a belief and desire that that belief and that desire cause them to act in that way. However, this is not a sufficient condition. Here is one of Davidson's many examples:

> [S]uppose, contrary to the legend, that Oedipus, for some dark oedipal reason, was hurrying along the road intent on killing his father, and, finding a surly old man blocking his way, killed him so he could (as he thought) get on with the main job. (Davidson, 1980b, p. 232)

Oedipus, in the example, kills his father, and his desire to kill his father causes him to do so—but, as Davidson says, we 'could not say … that his reason in killing the old man was to kill his

[33]. This is an aspect of Davidson's 'anomalous monism', the view that every mental event is identical to some physical event but that psychological laws are not reducible to physical laws.

[34]. Davidson's view is complex and I am necessarily leaving out a lot of detail, not least because his account of interpretation and psychological explanation is arguably inextricable from his philosophy of language and theories of truth, meaning and translation. Different aspects of Davidson's theory are developed in (Davidson, 1991) and many of the essays in (Davidson, 1980b, 2001), and elsewhere.

father' (Davidson, 1980b, p. 232). It seems that we need to specify more precisely the *way* in which beliefs and desire need to cause actions in order to rationalise them. As Davidson says, the desire and belief

> must cause [the action] in the right way, perhaps through a chain or process of reasoning that meets standards of rationality. I do not see how the right sort of causal process can be distinguished without, among other things, giving an account of how a decision is reached in the light of conflicting evidence and conflicting desires. ... [This] cannot be done without using notions like evidence, or good reasons for believing, and these notions outrun those with which we began. (Davidson, 1980b, pp. 232–3)

As he later puts the thought, in his response to Richard Peters on the same essay:

> In the formulation of hypotheses and the reading of evidence, there is no way psychology can avoid consideration of the nature of rationality, of coherence and consistency. At one end of the spectrum, logic and rational decision theory are psychological theories from which the obviously empirical has been drained. At the other end, there is some form of behaviourism better imagined than described from which all taint of the normative has been extracted. Psychology, if it deals with propositional attitudes, hovers in between. (Davidson, 1980b, p. 241)

The key idea is that in interpreting an agent, the way in which we update our theory of the agent's mind is sensitive to two different kinds of constraints which can in principle pull in different directions. On the one hand, there are the ordinary empirical considerations that constrain all scientific theories. The theory must fit the phenomena. On the other hand, though, in interpreting an agent in psychological terms we are also constrained by an ideal of rationality: we aim, necessarily, to understand them as approximating as closely as possible to ideal rationality. It is because this is a fundamentally different kind of constraint that psychological laws cannot be reduced to exceptionless causal laws. Given any psychophysical generalisation connecting psychological and physical predicates, we could never be sure that the generalisation is strictly true and projectible, because it would be in principle possible that we might find someone whose actions would be best rationalised by ascribing to them an attitude that the psychophysical generalisation would predict that they do not have.[35]

Functionalism and rationality

One kind of functionalist response to the Davidsonian argument would be to insist that considerations of rationality simply do not play the kind of role in ascriptions of mental states that Davidson claims they do.[36] An alternative, more congenial to the present investigation, is to say that while rationality does have a constitutive force in our ascriptions of mental states, this role can be captured by a functionalist account, because the relevant constraints of rationality can be satisfied by a physical system. One functionalist who develops an account of this kind is Brian Loar. Loar argues that constraints of rationality are partly constitutive of our concepts of mental states such as belief and desire, so that it is '*a priori* that if certain states are to be counted as beliefs and desires they must satisfy the

[35]. See (Loar, 1981, p. 20ff.) for an especially clear and concise presentation of this argument.
[36]. See, for instance, (Rey, 2007).

constraints of rationality' (Loar, 1981, pp. 23–4). However, this does not, Loar argues, rule out the possibility that certain physical states, in virtue of the structure of counterfactual relations between them, might contingently meet those requirements. Here, for example, is one of Loar's constraints:

> For physical state-types x and y to be related as the theory says the belief that p & q and the belief that ~p are counterfactually related is, in part, for it to be the case that if x were to occur then y would not occur. (Loar, 1981, p. 23)

Loar's suggestion is that by including enough such constraints in our theory of mental states, we can capture the constitutive role of rationality in a functionalist-friendly way.

Part of Loar's account is a list of constraints of 'minimal rationality' on the co-occurrence of beliefs, meant to capture the idea that we will tend not to ascribe to agents combinations of beliefs that are inconsistent in virtue of their logical form. As Loar puts it,

> The rationality constraints generate a vast network of such counterfactual relations among physical states, ultimately with the effect of describing a system of physical state types whose counterfactual interrelations mirror the relevant logical relations among beliefs and desires. (Loar, 1981, p. 23)

The only 'rationality constraints' Loar actually presents concern the co-occurrence of beliefs. However, he does also suggest that further principles will account for what agents tend to believe or want in given external circumstances. And, crucially, he also says something about how beliefs and desires interact to cause actions. Loar takes action-explanation to be captured by a certain kind of instrumental 'practical syllogism'—

> A desires that *q*;
>
> A believes that if *p* then A's doing [*V*] will lead to *q*;
>
> A believes that *p*.

—the conclusion of which is that A *V*-s. So, for instance, if Liz goes to Hereford because she thinks they have Jerseys there, what's going on is something like this:

1. Liz wants a Jersey;
2. Liz believes that if they have Jerseys in Hereford then her going to Hereford will lead to her getting a Jersey;
3. Liz believes that they have Jerseys in Hereford.

Liz's going to Hereford is caused by the interaction of these states and we understand it as such. As with Loar's constraints on co-occurrence of beliefs, the idea is that these mental states will tend to interact in this kind of way in virtue of their logical form. Of course, Loar's theory cannot give a full account of action explanation just by adding this model of practical inference onto the constraints of belief; there will also have to be generalisations about the sorts of things that humans in general, and perhaps humans of particular sorts, typically want in given circumstances, some generalisation to the effect that people generally have true

instrumental beliefs about how to achieve the things they want, and generalisations to the effect that people typically believe things that they are in a position to perceptually ascertain.

We might worry that an account like Loar's would threaten the kind of conception of rationalisation that I have been articulating. McDowell (1985), for instance, inspired by Davidson, argues that the 'rationality constraints' that Loar presents do not constitute a functionalist codification of the principles of rationality at play in rationalising explanations, and they do not, as Loar takes them to, show that such a thing is possible. Even if Loar's constraints do correspond well enough to genuine rational principles, McDowell argues, the fact that *some* such principles can be mirrored in purely causal generalisations fails to show that our whole concept of rationality can be captured in this way. And if the concept of rationality cannot be captured in this way, McDowell argues, the Loar-style functionalist account does not capture the idea of a person's doing the rational thing *because* it is rational.

Because Loar illustrates his view mainly with constraints on belief, McDowell's criticism is focused on theoretical rationality, and the point he presses is that Loar's account fails to capture 'the general normative notion of deductive consequence'—of 'what, in general, follows from what' (McDowell, 1998b). This general notion cannot, McDowell argues, be reflected in a set of principles statable in physical vocabulary, and because of this, a functionalist account makes unavailable

> a mode of understanding in which one finds a belief intelligible on the basis of its following deductively ... from other beliefs that one knows the believer holds. (McDowell, 1998b, p. 329)

It is not clear that McDowell's argument is successful. McDowell places great emphasis, for instance, on the holism of our conception of rationality. Since the kind of psychological theory to which the functionalist appeals is also holistic, though, it is unclear why this should be a problem. The key to the argument has to be the Davidsonian point about the way in which the constitutive ideal of rationality means that no psychophysical law could be strict and projectible. Both this argument and its relevance to functionalism, however, are complex and controversial.[37]

Another matter that is not entirely clear is the relation between a functionalist theory of mental states and a conception of what kinds of explanation and understanding our concepts of those states can figure in. We might, for instance, think that while we can derive implicit definitions of mental states by Ramsifying a 'theory' consisting of 'platitudes' about what a person will tend to do given that they are in certain kinds of mental states, there is nonetheless a kind of understanding of a person's actions that cannot be provided merely by such platitudes: namely, essentially first-personal understanding. This kind of understanding is possible because both we and the person who is the object of our understanding are self-conscious and act self-consciously. While functionalism as a theory of the nature of mental states might sit comfortably with the idea that psychological explanation is fundamentally

[37]. See (Yalowitz, 1997, 2014) for helpful discussion.

no different from any other kind of causal explanation, it is not clear that it *necessitates* that view.[38]

Even if the McDowell-style argument against functionalism is unsuccessful, Loar's specific account of how common-sense psychology involves constraints of rationality does not seem to be quite adequate. His minimal rationality constraints are arguably too minimal, and I think he fails to appreciate the significance of considerations of rationality in the explanation of action. However, if he is right that constraints of rationality can be worked into a functionalist theory, then perhaps functionalism is not inconsistent with the idea of rationalisation outlined in this chapter. If McDowell is right, on the other hand, then we need not worry about functionalism at all, because it is false. The general point I want to make here is just that while a crude analytic functionalism might seem to suggest a picture of psychological explanation that is at odds with the notion of rationalisation, this very fact gives us reason to either insist that an analytic functionalist account work with a conception of common-sense psychology that is sophisticated enough to accommodate the special character of rationalisation. If this cannot be done, we have reason to reject analytic functionalism.

[38]. (Bealer, 1997) argues that self-consciousness itself poses an insurmountable problem for reductive forms of functionalism, but his argument is no less contentious than Davidson's. For critical responses, see (Båve, 2017; McCullagh, 2000; Tooley, 2001).

Chapter 2

Responding to How Things Stand

2.1 Two kinds of rationalisation

I have so far emphasised two kinds of explanation of a person's action. A perspectival rationalisation explains an action by attributing a belief to the agent. I suggested in the previous chapter that when we understand an action on the basis of a perspectival rationalisation, our understanding is based on our grasp of (objective, worldly) reasons: what we come to appreciate, when we learn what the agent believed, is how, had things been as they took them to be, there would have been some reason to do what they did. I have also noted, but have said less about, worldly rationalisation, in which an action is explained by a worldly fact that was a reason for the agent to do what they did. Worldly rationalisation seems in a way to be more straightforward: the agent simply recognises a reason to do something and does it for that reason, and we understand their action just in seeing that there was a good reason for them to do what they did. We might, however, suspect a certain hidden complexity behind the apparently simple form of any worldly rationalisation. Specifically, we might think that every worldly rationalisation is to be properly understood in terms of a corresponding perspectival rationalisation. On this view, simply stating a fact that was the agent's reason for doing what they did is a convenient way to simultaneously *justify* the action in terms of a worldly reason and, at the same time, indicate a relevant belief of the agent's, where it is this belief that truly explains the action. If this is right, then when an action is explained by giving a worldly rationalisation, we come to understand the action by inferring a corresponding perspectival rationalisation. Or at least, the worldly character of the rationalisation does not add anything of psychologically explanatory significance to something that could be given in non-world-involving terms. For the purposes of understanding an action from the agent's point of view, whether their beliefs are true or false simply makes no difference.

An idea along these lines is nicely expressed by Davidson:

> Straight description of an intended result often explains an action better than stating that the result was intended or desired. "It will soothe your nerves" explains why I pour you a shot as efficiently as "I want to do something to soothe your nerves", since the first in the context of explanation implies the second; but the first does better, because, if it is true, the facts will justify my choice of action. Because justifying and explaining an action so often go hand in hand, we frequently indicate the primary reason[39] for an action by making a claim which, if true, would also verify, vindicate, or support the relevant belief or attitude of the agent. "I knew I ought to return it", "The paper said it was going to snow", "You stepped on my toes", all, in appropriate reason-giving contexts, perform this familiar dual function.
>
> The justifying role of a reason, given this interpretation, depends upon the explanatory role, but the converse does not hold. Your stepping on my toes neither explains nor justifies my stepping on your toes unless I believe you stepped on my toes, but the belief alone, true or false, explains my action. (Davidson, 1980b, p. 8)

It is worth considering, though, whether we *have* to view worldly rationalisation in this way. Might it not be the case that, at least sometimes, we fundamentally understand someone (perhaps ourselves) as responding to how things really are, and that there is thus a form of rationalisation that is irreducibly 'worldly' or factive? On this view, our understanding of an action as rational from the agent's perspective is not only based on our general understanding of worldly reasons, as I suggested in the last chapter; in some cases, it fundamentally involves recognising the action as having been taken because there was a reason to take it.

I will call such an insistence on the irreducibility of worldly rationalisation *factivism*. The Davidsonian view, that worldly rationalisations are fundamentally to be understood in terms of perspectival rationalisations, I will call the *perspectivalist* view. In this chapter I will consider two arguments for the perspectivalist view. Both are based on the asymmetric dependence of worldly rationalisations on perspectival rationalisations—the point, emphasised by Davidson in the quote above, that a worldly fact explains an action as its reason only if the agent believes that the fact obtains, whereas the belief can explain the action regardless whether it is true or not. However, the arguments exploit this idea in different ways. One of the two arguments is more compelling than the other, but it also leaves some room for an argument in support of the factivist view. I will outline one such argument, showing how factivism as applied to the rationalisation of perceptually-based *belief* makes possible an attractive account of the epistemology of perception. Finally, I will consider the prospects for generalising the conclusion of this argument to the practical case.

2.2 Individualism and psychological explanation

Factivism, as I am defining it, is the view that there are rationalising explanations whose psychologically explanatory role cannot be reduced to that of facts about the agent's perspective on the world. In other words, there are some actions or attitudes whose

[39]. For Davidson, the 'primary reason' for an action is a belief–desire pair that rationalises the action, not the kind of worldly fact that constitutes a reason on the present account.

intelligibility we can only fully grasp when we understand those actions or attitudes as responses to facts 'out there in the world'. This idea is inconsistent with *psychological individualism*, the view that the psychologically explanatory facts about an agent are fixed by how that agent is intrinsically or internally. Since the worldly facts do not in general depend on how the agent is intrinsically, psychological individualism implies that worldly rationalisations are genuinely psychologically explanatory only insofar as they carry information about facts that *do* depend, and depend solely, on how the agent is intrinsically. Since each worldly rationalisation seems to imply a corresponding perspectival rationalisation—if I go to Hereford because they have Jerseys then I go to Hereford because I think they have Jerseys—then the perspectivalist view seems the best account of how worldly rationalisations get to be explanatory, assuming, as the individualist presumably will, that an agent's beliefs are fixed by their intrinsic states.

Individualism imposes a strict constraint on what kinds of considerations can be genuinely psychologically explanatory. It is clear enough that individualism would rule out factivism if it were true. The question is whether we have any compelling reason to accept it. I will argue, against individualism, that there is good reason to think that facts depending not just on how things are intrinsically with the agent, but also on how things are in the agent's environment and the agent's relations to that environment, can be genuinely psychologically explanatory, and that there is no compelling reason to expect the explanatory force of these facts to reduce to that of facts fixed by the agent's intrinsic state. Hence we have good reason to reject individualism as a general principle, hence it fails to provide a sound basis for ruling out the factivist position.

2.2.1 The argument for individualism

Individualism of the sort that would threaten factivism might be motivated roughly as follows. Psychological explanations of a person's attitudes or behaviour are causal explanations of that person's attitudes or behaviour. As such, a psychological explanation imputes causal efficacy to the conditions mentioned in its *explanans*. The immediate causes of a person's behaviour and attitudes, however, are intrinsic conditions of that person. Since an agent's behaviour and attitudes are immediately caused by her intrinsic states, only differences in the agent's intrinsic condition make a causally explanatory difference to what she does. Features of the world 'out there' can, of course, affect what an agent does, but only mediately, via influencing the agent's intrinsic condition.

Operative here is the idea that psychological conditions are typed in terms of the kind of behaviour and attitudes that they are liable to explain. According to Jerry Fodor's (1991) version of the individualist argument, consideration of certain kinds of cases reveals that any putative difference in the behaviour of two intrinsically qualitatively identical agents will conceptually depend upon a corresponding difference in the relationally-individuated content of their psychological states. This means that the applicability of certain relational descriptions to an agent's behaviour depends on how things are in the agent's environment,

but, Fodor argues, this isn't a genuine causal difference, since at a more basic level of description the behaviour will be the same.

Fodor's argument is based on a certain familiar form of thought experiment, in which we are asked to imagine two qualitatively identical subjects in environments that differ in a specific kind of way. So: Jerry is here on Earth, while his 'twin', Gerry, is on Twin Earth, the only difference between Earth and Twin Earth being that on the latter, instead of water, they have twater, a substance indiscriminable from water but with a different molecular make-up. This style of example of course originates in arguments for semantic externalism, the idea that meanings are not 'in the head': if Jerry says 'Water is wet,' he seems to be referring to water, wheres if Gerry makes the same utterance, he will be referring to twater. Since this difference in meaning corresponds to no difference in the twins' intrinsic qualities, the meanings of their utterances cannot be fixed by their intrinsic qualities.[40] The same kind of argument can be extended to support externalism about psychological content, which makes a corresponding claim about what Jerry and Gerry believe (judge, suppose, and so on), rather than just the meanings of the words they utter.[41]

Fodor is happy to grant these externalist conclusions—he simply claims that the differences in meaning and content, the differences that are not fixed by the twins' intrinsic qualities, are not relevant for the purposes of psychological explanation. True, when Jerry wants water and thinks there is water in the kitchen, he goes to the kitchen to get water (not twater), and when Gerry wants twater and thinks there is twater in the kitchen, he goes to the kitchen to get twater (not water), so the difference in mental content corresponds to a difference in behaviour. However, this difference, Fodor argues, is neither genuinely causal nor genuinely psychologically explanatory. The difference is merely conceptual or logical, given the way that the content of an intention in action is conceptually fixed by the contents of the mental states that produce it. If we switched the twins' places, Jerry would still intend to get water and Gerry would still intend to get twater, but in a very real sense what each would actually *do* would be exactly what the other would have done in his place. There is a difference in how their behaviour can be properly described in intentional terms, but this difference is merely conceptual, so non-contingent, so not causally explanatory.

2.2.2 Explanatory proportionality

The similarity between Jerry's and Gerry's behaviour is certainly impressive, but given the contrived nature of the case it will do to consider other examples. In their original use in supporting externalism, the contrivance is perhaps less problematic: the point of Twin Earth cases there is to provide a counterexample to a universal claim. In Fodor's use, however, things are a bit more complicated. Here we are concerned with the causal-explanatory relevance of intrinsic versus extrinsic or relational conditions, and it is not clear that we can assess this just by considering a narrow range of cases in which both the intrinsic and most of the environmental conditions are fixed in this way. In this context, it is crucial to consider

[40]. See (Putnam, 1975).
[41]. (Burge, 1979, 1986; McGinn, 1977).

similarities across cases, in particular what happens when some specific condition is kept fixed while others are changed. Twin Earth cases fix intrinsic conditions and alter environmental conditions only in a very specific way. Other cross-case comparisons might not be so favourable to the individualist position.[42]

Suppose Jerry is thirsty and wants an *Americano*. He thinks the *Punt e Mes* is in the cabinet, so he goes to the cabinet. Jerry goes to the cabinet because he thinks that's where the *Punt e Mes* is. 'Going to the cabinet' is a relational description of Jerry's behaviour, in that whether some movement of Jerry's body is an action of going to the cabinet depends on where the cabinet is in relation to his body. A movement does not count as an action of going to the cabinet just in virtue of the contents of the mental states that cause it: to paraphrase a point made by Peacocke (1993), it is not a conceptual necessity that people with thoughts about cabinets produce behaviour that involves relations to cabinets. Jerry might have been in an unfamiliar environment where, although he believes that there is *Punt e Mes* in the cabinet, he has no idea, or perhaps a false belief about, where the cabinet is. However, although there is not this kind of conceptual necessity here, the fact that Jerry wants an *Americano* and thinks that there is *Punt e Mes* in the cabinet does support certain counterfactuals: it suggests that if the cabinet were in the lounge, he would go to the lounge, and that if it were in the basement, he would go to the basement. Of course, these counterfactuals will only be supported in this way given that certain other contingent conditions are met—Jerry knows where the cabinet is, the route to the cabinet is not booby-trapped, and so on—but these are conditions that will be met in ordinary contexts of explanation, and which we would in many contexts assume to be met without necessarily knowing much, if anything, about Jerry's intrinsic condition. We need to make such assumptions in order to get Fodor's case going, too: we assume that the twins know where the next room is and how to get there. Any psychological explanation needs such background conditions in order to get off the ground.

Seemingly, then, Jerry's belief that the *Punt e Mes* is in the cabinet stands to explain his going to the cabinet across a range of environmental conditions. Moreover, that belief can also explain that action across a variety of *intrinsic* conditions of Jerry. Since Jerry knows the location of the cabinet in each case but in each case that location is different, Jerry will presumably be different intrinsically in some way. This is reflected also in the fact that in each case the intrinsic, non-relational description of his movements will be different. We could add in further variety: Jerry could have found out about the location of the *Punt e Mes* in different ways, could take different means of approach to the cabinet, and so on. Most likely each of these differences will involve some difference in Jerry's intrinsic condition, but the ability of his belief to explain his going to the cabinet holds despite this variation in intrinsic condition. It seems, then, that the relational, environment-involving, descriptions of Jerry's mental states and actions allow for an explanatory generality that is not obviously afforded if we restrict ourselves to intrinsic descriptions of his state and behaviour.

[42]. The argument of this section owes much to (Peacocke, 1993; Stalnaker, 1989, 1990; Williamson, 2000; Yablo, 1992).

This by no means disproves the individualist theory. It is quite possible that, although there are differences in Jerry's intrinsic condition across the cases, there is some individualist-friendly state, some aspect of his intrinsic condition, that is constant and that explains his action, and that this state either just is his belief or is something like an intrinsic component of his belief. The intrinsic description of his action is different in each case, of course, but the individualist can insist that that results from the differences in his intrinsic condition, in particular the differences corresponding to the differences in his belief about where specifically the cabinet is. However, while it is conceivable that there is such a common intrinsic state that explains his actions across the cases, this is just a hypothesis, and it is a live question whether we have any reason to accept it apart from the fact that it would save individualism.

I do not mean to argue that what psychologically explains a person's actions in no way depends on that person's intrinsic condition, such as their brain states. Of course Jerry's thinking that there is *Punt e Mes* in the cabinet is not consistent with his being just any old way intrinsically. Similarly, if Jerry is to go to the cabinet, his body must move in the right 'intrinsic' way. It seems eminently plausible that, if Jerry goes to the cabinet because he thinks that that is where the *Punt e Mes* is, then the intrinsic movements of his body involved in his going to the cabinet must have been caused, at least in part, by the intrinsic conditions involved in his believing that the *Punt e Mes* is in the cabinet. This of course fits nicely with the idea, to which the Twin Earth examples appeal, that intrinsic duplicates will behave in intrinsically similar ways. However, it does not obviously imply that believing is a purely intrinsic condition or that explanations in terms of the agent's intrinsic condition can match the explanatory generality of explanations in terms of belief. What it says is that in any instance of Jerry's going to the cabinet because he thinks there is *Punt e Mes* in it, there is some intrinsic realiser of his believing and some intrinsic realiser of his action, and the former causes the latter. Presumably the intrinsic state also explains the occurrence of the intrinsic event, and does so in virtue of its being a state of a general kind where states of that kind tend to cause events of the kind that the intrinsic realiser of the action is.

We can grant, then, that whenever a belief explains an action, there are some intrinsic realiser states and some intrinsic realiser movements that stand in this kind of explanatory relation. However, the individualist needs to make a much stronger claim: that where there is an explanatorily relevant commonality in psychological state, as there is across the cases of Jerry's going to the cabinet because he thinks the *Punt e Mes* is in it, that state is, across the relevant cases, realised by intrinsic states of a common kind. That is, the psychologically explanatory generality must be matched by a generality covering intrinsic conditions. As Fodor himself effectively argued in a much earlier paper (Fodor, 1974), the moderate concession I have made about psychological states having intrinsic realisers in no way entails this stronger individualist claim. If there is some individualistic predicate that Jerry satisfies in every case of his believing that the *Punt e Mes* is in the cabinet, we have been given no reason to expect it to be anything other than, in the earlier Fodor's phrase, 'wildly disjunctive'—and, crucially, open-endedly disjunctive, the only reason for putting the

disjuncts together being that they are the possible realisers of Jerry's believing that the *Punt e Mes* is in the cabinet. The same applies to the intrinsic realisers of his action.

Individualism says that unless there is some intrinsic condition in common across the cases, then the apparent explanatory generality of the fact that Jerry believes that the *Punt e Mes* is in the cabinet is illusory. Other than an insistence on individualism, though, we have been given no reason to believe that there is such an intrinsic commonality. Yet we do have reason to believe that there are natural psychological predicates that apply across the cases: '… thinks that the *Punt e Mes* is in the cabinet' and '… goes to the cabinet'. We exploit the broad applicability of these predicates in our rationalising explanation, and their generality makes them useful and explanatorily powerful. Such predicates afford ways of understanding people's behaviour that do not presuppose any detailed knowledge of their internal states or intrinsic bodily movements. Given how often we have to make sense of each other in ignorance of those details, this is quite a blessing.

Fodor's *a priori* individualism, then, fails to show that external conditions cannot be genuinely and irreducibly psychologically explanatory. Nothing I have said rules out the possibility of supporting individualism on more empirical grounds, of course. However, assessing the empirical support for such a view will involve, in part, seeing how well it accounts for ordinary psychological explanation. For now at least, factivism remains a live option: for all we have seen, the mere externality of worldly reasons does not necessarily mean that they could not play an irreducible role in psychological explanation.

2.3 Davidson's argument

I have argued that the individualist argument against factivism—the view that worldly rationalisations, explanations of actions that appeal directly to the objective, 'worldly' reasons for which the agent acted, are not reducible to explanations in terms of the agent's non-factive mental states—fails. My rejection of that argument was, in essence, based on the fact that individualism is not itself well enough motivated, and in particular that it flouts good principles of explanatory proportionality. Without compelling motivation, the restrictions the individualist imposes on what kinds of considerations can be genuinely psychologically explanatory appear arbitrary.

However, this does not simply clear the way for factivism, and a view about psychological explanation as controversial as individualism is not the only basis for reasonable scepticism about the irreducibility of worldly rationalisation. A more modest case can be made for the perspectivalist view, and in fact it can be made on the basis of principles of explanatory proportionality of much the same kind as those we used to call individualism into question. The perspectivalist theorist holds that when we explain a person's action by simply stating a fact that was a reason for them to act as they did, as in 'Liz went to Hereford because they have Jerseys there', this rationalises the action only by implying something about what the agent believed to be the case. In other words, 'Liz went to Hereford because they have Jerseys there' counts as a rationalisation because, as well as telling us that they do in

fact have Jerseys in Hereford, it also implies that Liz believed that they do, and the explanation makes first-personal sense of Liz's action just because and just insofar as it tells us that she acted on that belief.

The first point to note is that it is very plausible that any true worldly rationalisation implies the truth of a corresponding perspectival rationalisation. If Liz went to Hereford because they have Jerseys there, where that fact is the reason for which Liz acted, then it seems to follow that Liz went to Hereford because she thought they have Jerseys there. Both perspectivalist and factivist theorists will agree on this point. This simple observation suggests a straightforward argument for the perspectivalist view. Liz's belief alone is enough to make sense of her action, whether or not that belief is correct. In light of this, it is unclear what the *truth* of the belief could be adding. Precisely this form of argument is suggested by the passage from Davidson that I quoted at the start of this chapter. Davidson begins by explaining why, although rationalisation is essentially perspectival, we often give rationalisng explanations that are not in explicitly perspectival form:

> Straight description of an intended result often explains an action better than stating that the result was intended or desired. 'It will soothe your nerves' explains why I pour you a shot as efficiently as 'I want to do something to soothe your nerves', since the first in the context of explanation implies the second; but the first does better, because, if it is true, the facts will justify my choice of action. (Davidson, 1980a, p. 8)

Although Davidson is here focused on explanations in terms of aims, intentions and desires, the thought clearly applies equally well to those in terms of reasons and beliefs. Davidson's thought is that because when we seek to explain actions we are also at the same time interested in whether those actions are justified, it is simply more efficient, when possible, to give what I am calling a worldly rationalisation: worldly rationalisations, unlike perspectival rationalisations, perform a 'dual function', telling us both why a person did what they did and also what reason there actually was for them to do it. Davidson then presents his simple argument for the perspectivalist view:

> The justifying role of a reason, given this interpretation, depends upon the explanatory role, but the converse does not hold. Your stepping on my toes neither explains nor justifies my stepping on your toes unless I believe you stepped on my toes, but the belief alone, true or false, explains my action (Davidson, 1980a, p. 8)

2.3.1 A wrinkle: worldly rationalisation and knowledge

There is a slight complication here that the factivist will be quick to point out, which is that a worldly rationalisation does, arguably at least, give us more information about the agent's psychology than the corresponding perspectival rationalisation alone. The quote from Davidson, in ascribing to worldly rationalisations a 'dual function', might be understood to say that all a worldly rationalisation does is to (a) state a fact about the world and (b) imply that the agent believes that that fact obtains. This would be a mistake: 'A *V*-ed because *p*' is

not equivalent to '*p* and A *V*-ed because A believed that *p*'. There must be some real explanatory connection between its actually being the case that *p* and A's *V*-ing. Suppose that, although it is true that they have Jerseys in Hereford, Liz only thinks this because a colleague told her that they have Jerseys in Hertford and she later mixed the two towns up. Here it seems that, while it is true that they have Jerseys in Hereford, and it is true that Liz goes to Hereford because she thinks they have Jerseys there, it is *not* the case that Liz goes to Hereford *because* they have Jerseys there: her belief is not sensitive in the right way to that fact's really obtaining, and so neither is her action. The availability of a worldly rationalisation seems to require not just a lining up of the agent's perspective with how the world really is, but a certain kind of connection between the world and the agent's perspective. If someone acts on a truth but only by luck, the truth does not explain their acting as they do.

A number of authors, considering similar kinds of cases, have argued that, if someone is to act *for* a worldly reason, and so for a worldly rationalisation in terms of that reason to be available to explain their action, the person must *know* the reason in question.[43] Worldly rationalisation seems to be unavailable in just the kinds of cases in which the agent fails to know the fact in question: where their belief is mistaken, or unreasonable, or 'Gettiered'. The best explanation for this, these authors argue, is that in order to act for some worldly reason, you need to know that reason. If these authors are right, then 'A *V*-ed because *p*' implies not just that A *V*-ed because she thought that *p*, but also that A knew that *p*. Only on an extreme externalism about knowledge would this imply nothing of psychological significance beyond the fact that the agent believed that *p*. If, for example, knowing that *p* involves believing that *p* rationally or with adequate justification, then a worldly rationalisation tells us not just that the agent acted on a belief that was, as it happens, true, but also that they held that belief rationally or with adequate justification, and (again, unless an extreme form of externalism about these notions is true) then this tells us more about the psychological conditions leading to the action than the quote from Davidson suggests.

However, if there is an interesting perspectivalist position, it ought not to be challenged by this observation alone. The perspectivalist view is opposed to factivism, and the factivist idea is that worldly rationalisation is a special form of rationalisation, that there is something important about a worldly reason itself making a response to it intelligible from the agent's point of view. In some cases, the factivist says, understanding an action first-personally involves recognising that the agent had a certain fact *in* view and acted because that fact obtains. To reject this, the perspectivalist theorist need not claim that there is no relevant psychologically explanatory information that a worldly rationalisation provides as against a perspectival rationalisation; the perspectivalist claim is just that whatever psychologically explanatory information a worldly rationalisation does provide, the agent's having got things right in the particular case is not essential to the way that that information makes sense of that agent's doings. It is certainly not obvious that the extra information provided by a worldly rationalisation has this character.

[43]. See in particular (Hornsby, 2008; Hyman, 1999; McDowell, 2013)

Later in this chapter I will consider whether an argument for factivism might in fact be mounted on the basis of the role of knowledge in making worldly rationalisations possible. First, though, I want to discuss in more detail the Davidsonian argument for the primacy of perspectival explanation.

2.3.2 Perspectival rationalisations as proportionate explanations

The Davidsonian argument for the perspectivalist approach appeals to the generality of perspectival rationalisation. There appear to be relevant similarities between cases—similarities in the actions of agents and the psychological precursors of these actions—that perspectival rationalisations capture but that worldly rationalisations do not. We can group together certain cases as having something significant in common, for example: (i) you stepped on my toes and I am aware of this, (ii) you stepped on my toes but my belief is 'Gettiered', say because I did not feel you stepping on my toes but a very short person, who I did not notice, stepped on my other foot at the same time, or (iii) you did not step on my toes and I merely think that you did (perhaps because of the short person again). In any of the cases (i)–(iii), we could explain my stepping on your toes by saying that I thought you stepped on my toes, but only in (i) could we say that I stepped on your toes because you stepped on my toes.

The thought here is the simple one that: first, if the fact that *p* explains the agent's *V*-ing by being a reason for which the agent *V*-ed, then the agent's *V*-ing in the same case can be explained merely by their believing that *p*; second, their believing that *p* would explain their *V*-ing even if it were not the case that *p*, and would do so in the same way, namely by making the action intelligible as an exercise of the agent's rationality. This point does not rest on any claims about intrinsic proximal causes; it is simply the point that when we can give a worldly rationalisation, the agent's having a corresponding belief is, but the fact's obtaining is not, necessary for making the agent's action rationally intelligible. This is taken as evidence for the explanatory priority of facts about the agent's non-factive mental states. To assess the argument, we need to consider why it should be so taken.

One way to sharpen the argument is by appealing to the kinds of considerations I appealed to above in assessing the individualist argument against factivism. There, considerations of explanatory generality were used to reject a methodological constraint that would rule world-involving conditions psychologically irrelevant, the argument being that world-involving conditions are, plausibly, better proportioned to the outcome (that is, the person's performing the action that they do) than any purely intrinsic conditions. Here, the greater explanatory generality of a non-factive psychological condition, belief in the existence of a worldly reason, seems to make it better proportioned to the outcome than is the actual existence of such a reason.

This line of thought about explanatory relevance is closely related to concerns about causal relevance—perhaps unsurprisingly, given that rationalisations are at least in some sense causal explanations—and as such the argument can usefully be formalised using some

machinery developed to address issues of causal relevance. Stephen Yablo (2003) articulates what he calls the *proportionality theory of causal relevance*. The theory is inspired by counterfactual theories of causal relevance, but resolves problems such theories face concerning the counterfactual relevance of properties that, when we consider *causal* relevance, seem unnecessarily weak or strong. Yablo rules out such cases as follows:

- A property P of *x* is egregiously weak (relative to effect *y*) iff some more natural stronger property of *x* is better proportioned to *y* than P is.
- A property P of *x* is egregiously strong (relative to effect *y*) iff some as natural weaker property of *x* is better proportioned to *y* than P is.

Proportionality is assessed in terms of counterfactuals:

- Q– is better proportioned to *y* than Q+ iff *y* would still have occurred, had *x* possessed Q– but not Q+.
- Q– is worse proportioned to *y* than Q+ iff *y* would not have occurred, had *x* possessed Q– but not Q+.

With these terms so defined, we can state the proportionality theory of causal relevance:

- A property P of *x* is causally relevant to effect *y* iff
  a) had *x* lacked P, *y* would not have occurred
  b) P is not egregiously weak or strong.[44]

Since Yablo's theory is articulated in terms of properties, if we want to use it to compare the causal relevance of reasons and beliefs, we will need, somewhat infelicitously but not too problematically, to represent the presence of a worldly reason as a property of the agent, such as *being in a world in which* p. How does such a property compare with the agent's belief that *p* for causal relevance, in a case in which the agent's *V*-ing can be given a worldly rationalisation, 'A *V*-ed because *p*'?

Where A *V*s because *p*, the agent's being in a *p*-world does appear to be causally relevant. Where a worldly rationalisation can be given, this suggests that, had it not been the case that *p* (had A not been in a *p*-world), A would not have *V*-ed. Is A's being in a *p*-world egregiously weak or strong? This will depend on the case and what we fill in for '*p*', but it is not as a general rule the case that A's being in a *p*-world is egregiously strong relative to A's believing that *p*. As a general rule, neither property will be weaker or stronger than the other, since neither, as a rule, entails the other. Similar considerations suggest that A's believing *p* is also causally relevant.

However, the point of the Davidsonian argument, as I think we should understand it, was never meant to be that worldly reasons are never causally or explanatorily relevant to actions. Rather, it is that when they are relevant to an action, this is only in virtue of their causal-explanatory relevance to the agent's having a belief that is causally explanatorily relevant to the action. The Davidsonian argument, then, is not best put by comparing the causal relevance of the existence of a worldly reason with that of the agent's believing in the existence of that reason, but rather by comparing the relevance of the agent's having that belief with that of the agent's being such that she can act for the reason in question. Call this

[44]. (Yablo, 2003, p. 342 apparent error corrected).

latter property, of an agent's being such that her actions can be given worldly rationalisations in terms of the fact that *p*, the agent's being *aware* that *p*.[45] The argument for the perspectivalist view is that, when we compare it to the belief that *p*, awareness that *p* appears to be egregiously strong. In the kind of case we are considering, A's being aware that *p* might meet the first condition for causal relevance: it will often be plausible in such cases to say that had the agent not been aware that *p*, she would not have *V*-ed. Such cases will be those in which, had she not been aware that *p*, she would not have believed that *p*. However, awareness, the Davidsonian argument goes, seems to fail the second condition: it is egregiously strong. This is because believing that *p* seems to be at least as natural a property as being aware that *p*, but is better proportioned to the result of A's V-ing: A would still have V-ed, had she believed that *p* but not been aware that *p* (because her belief was false, Gettiered, or whatever).[46] So awareness is not causally relevant. The same reasoning that threatens the causal relevance of awareness seems also to threaten its rational-explanatory relevance. Not only would A still have *V*-ed had she believed *p* but not been aware that *p*; but, her action would still have been rational, would still have been intelligible from her point of view, and in the same kind of way. So whatever awareness adds, it seems to be extraneous for the purposes of rationalising A's action.

Note that this argument does not depend on the idea that awareness is *less* natural a condition than belief, as we might think if we were of the view that awareness is reducible in terms of belief. Suppose, for example, that, as some authors have argued, being aware that *p*, in the present sense, just is knowing that *p*, and suppose that we accept Williamson's (2000) arguments for the irreducibility of knowledge. Even if we grant that knowledge is an irreducible mental state, just as natural as belief, it is still true that the agent's belief is better proportioned to her action than her knowledge is, so knowledge is still egregiously strong.

To avoid the conclusion that awareness of a reason is explanatorily 'screened off' by belief, it seems that the factivist will have to claim that awareness is *more* natural than belief. To see the kind of structure the factivist theorist might be positing here, consider the following example. Suppose we are in case (i): I step on your toes because you stepped on mine. Suppose also that I hate spiders and try to kill them whenever I see them. Now consider the proposition that either I noticed that you stepped on my toes or there is a noticeable spider on your foot. This disjunction is better correlated with my stepping on your foot than is the fact that you stepped on my foot. Let's use the following assignment:

[45]. According to Hyman, Hornsby and Williamson, being aware that *p* in this sense is just knowing that *p*. I do not want to commit to this equation of awareness and knowledge, for reasons that will become apparent later in the chapter.

[46]. This is complicated by cases where A's being aware that *p* is relevant to her ability actually to V. Given the possibility of such cases, the counterfactual does not hold as a general rule. However, the proponent of the Davidsonian argument can respond that, when we are engaged in rationalisation, we are primarily concerned with understanding the agent's intentions in acting, and the cases of failure due to lack of awareness are failures of execution rather than cases in which the agent acts with a different intention. Assessing the adequacy of this response would take me too far afield, but see (Gibbons, 2001).

- *p* : You stepped on my toes;
- *q* : I noticed that you stepped on my toes;
- *r* : there is a noticeable spider on your foot;
- *V*-ing : stepping on your toes.

Call the case in which *r* and not-*p* case (iv). Now, in both cases (i) and (iv), the following conditions hold:

1) *q* or *r*
2) Had it been the case that (*p* and not-(*q* or *r*)), I would not have *V*'d.
3) Had it been the case that ((*q* or *r*) and not-*p*), I would have *V*'d.

In case (i), (2) is true simply because had I not noticed you stepping on my toes, I would not have stepped on yours—the spider has nothing to do with it. (3) is true because had it not been the case that you stepped on my toes, I could not have noticed that you stepped on my toes, so the protasis implies that there is a very noticeable spider on your foot, so I would have stepped on your foot, but for entirely different reasons. Similarly, in case (iv), (2) is true because had the spider not been there, I would not have stepped on your toes—the possibility of your stepping on my toes has nothing to do with it—and (3) is true simply because it is the case that there is a very noticeable spider on your foot and I did step on your toes. I take it to be obvious that the fact that (2) and (3) hold in (i) does not show that 'I stepped on your toes because either I noticed that you stepped on my toes or there was a very noticeable spider on your foot' gives a more fundamental explanation of my action than does 'I stepped on your toes because you stepped on mine'. It does not, because the disjunction (*q* or *r*) is less natural than *p*, and there are good reasons to think as much, independent of those concerning proportionality to the outcome of my stepping on your toes.

The problem for the factivist is that believing seems to be a much more natural condition than the kind of disjunctive condition just discussed. Short of independent reasons for thinking belief less natural than awareness, or for rejecting one or other of the argument's assumptions, the argument from explanatory proportionality seems to support the perspectivalist view. However, whereas the individualist argument, had it been successful, would have ruled out *a priori* the explanatory significance of awareness of a reason, the proportionality-based argument leaves room for the proponent of the factivist view to say more. We might well construe the disagreement between perspectivalist and factivist theories as in effect a disagreement about whether being aware of a fact is a more natural condition than believing that that fact obtains. A positive argument for factivism will potentially weigh against the proportionality-based argument. If there is no compelling argument for factivism, the proportionality-based argument should be taken to provide good grounds for treating the perspectivalist view as in effect the default position. If a compelling argument for factivism can be articulated, though, it is not obvious that the proportionality-based argument should be taken to undermine it. The proportionality considerations do not show factivism to be *impossible*. We should, then, consider possible arguments for factivism.

2.4 Factivism and disjunctivism

The factivist position on the relation between worldly and perspectival rationalisation is in some respects akin to disjunctivist views in the philosophy of perception and perceptual knowledge. Much as the disjunctivist denies that all perceptual experience is to be understood in terms of an element common between veridical and non-veridical (illusory or hallucinatory) experience, the factivist denies that worldly and perspectival rationalisation are to be understood in terms of a common element. There is in the case of perceptual experience a similar asymmetrical dependence of a world-involving condition (acting because *p*; seeing that *p*) on a non-world-involving condition (acting because one believes that *p*; its visually seeming to one that *p*), but disjunctivists deny that this shows that the latter condition is more fundamental.

However, there are different motivations for disjunctivism, which lead in turn to different kinds of disjunctivist theory. One motivation is to defend 'naïve realism' about object-perception, the view that the experience one has when one sees a mind-independent object is partly constituted by that object. That position seems to be threatened by the observation that it is possible to have an hallucinatory experience that is subjectively indistinguishable from an experience of seeing a given object, and that one can have such an experience in the absence of any such object. The initial disjunctivist response is that we cannot simply assume, because of the way we group these together as 'experiences', that they are occurrences of the same fundamental kind. This kind 'experience as of seeing …' could be essentially disjunctive, such that to have a visual experience as of an object *O* could be either to see *O* or to merely seem to see *O*.[47] The challenge for this kind of disjunctivist is how to respond to the *causal argument from hallucination*, which purports to show that hallucinations and veridical perceptions are in fact of the same fundamental kind. There are responses to this argument, but they are controversial.[48] There is no need for us to take a position on these issues concerning the conscious character of perceptual experience, though, because the concerns are largely unique to that subject matter. The similarity between factivism and disjunctivism about the phenomenal character of perceptual experience is suggestive, but relatively superficial.

A stronger parallel can be found if we consider forms of disjunctivism that are motivated primarily by epistemological concerns. Here a disjunctivist account of perceptual *knowledge* is endorsed in order to block a threat to the possibility of such knowledge, a threat that comes from the *argument from illusion*. As John McDowell, probably the most influential proponent of this second form of disjunctivism, summarises that argument, it says that

> since there can be deceptive cases experientially indistinguishable from non-deceptive cases, one's experiential intake … must be the same in both kinds of case. In a deceptive case, one's experiential intake must *ex hypothesi* fall

[47]. See (Hinton, 1973).

[48]. See (Robinson, 1985, 1994 for the causal argument) (Burge, 2005; Martin, 2004 for an influential response, 2006 for responses to some of these objections; Siegel, 2004; Sturgeon, 1998 for objections to Martin's approach).

> short of the fact itself, in the sense of being consistent with there being no such fact. So that must be true, according to the argument, in a non-deceptive case too. (McDowell, 1998a, p. 386)

'Experiential intake' here means something like: what evidence perception provides us with, or what facts perception makes reflectively available. The issue of experiential intake here, then, is different from the question about the objects of perceptual experience with which the naïve realist disjunctivist was concerned. From the conclusion that our experiential intake falls short of the (worldly) fact itself, the argument from illusion continues to say that, when we acquire knowledge through our capacity to tell how things are by looking,

> we have to conceive the basis [of our knowledge] as a highest common factor of what is available to experience in the deceptive and non-deceptive cases alike, and hence as something that is at best a defeasible ground for knowledge … . (McDowell, 1998a, p. 386)

'Defeasible' here means non-conclusive: what is available to experience, if we accept the argument from illusion, is a defeasible ground for knowledge in that it is consistent with the evidence with which experience provides us that things are not as they seem. The McDowellian disjunctivist's concern is that, if what seeing that *p* makes us aware of were something that falls short of its actually being the case that *p*, then acquiring knowledge of the external world would require us to make an inference from the grounds that experience does provide. This inference would need to be supported by some hypothesis connecting how things seem to how things actually are, such as the hypothesis that our experiences are generally reliable. Belief in such a hypothesis, however, could only be justified by knowledge gained from experience. Hence, the 'highest common factor' view of experience leads to scepticism. To avoid the sceptical conclusion, McDowell argues, we must conceive of experience as providing grounds in the good case that it cannot provide in the bad case, namely *factive* grounds, grounds that entail the truth of our perceptual beliefs. That such grounds are only available in the good case is what makes the account disjunctivist: experience either puts us in touch with the facts or merely seems to.

This kind of disjunctivism clearly bears a much closer relation to our present concerns about action explanation. It is concerned with justification, the agent's point of view, and rational connections with worldly facts. Both the factivist and the McDowellian disjunctivist hold that making proper sense of a certain aspect of our rational psychology requires us to attribute, in some cases, a kind of connection between a fact and a rational response which cannot be fully understood in terms of what it shares with its non-factive analogue. This is quite different from the disjunctivism concerned with object perception, which is stimulated by problems quite proprietary to its metaphysical picture of the nature of perceptual experience. Because of this, it is not clear that the epistemological disjunctivist, or the factivist, must reject the idea that there is a relevant psychological (experiential) element in common between the good and bad cases.[49] Factivism may, then, be consistent with a merely 'non-conjunctivist' position.[50] On this approach, we simply reject the perspectivalist

[49]. (Byrne & Logue, 2008; Snowdon, 2005)
[50]. See (Williamson, 2000, pp. 44–8).

theorist's claim about the priority of perspectival rationalisation, and say that at least sometimes, when an agent A *V*-s because *p*, the content of this explanation cannot be (non-trivially) captured by the conjunction of A's *V*-ing because she thought that *p* with further conditions. We need not add, as the object-perception disjunctivist might regarding the case that concerns them, that there is no distinctive element, picked out by perspectival rationalisation, that is common across cases in which worldly rationalisation is available and those in which it is not.[51]

The question now is what exactly it is that the factivist thinks we need to posit this structure in order to explain. We saw how the McDowellian disjunctivist about perceptual knowledge is motivated by the need to explain how perceptual knowledge is possible. What plays the corresponding role in motivating factivism about the rationalisation of action?

2.4.1 Roessler's argument

One concern about the perspectivalist view might be that it makes too sharp a distinction between normative and explanatory questions about an action. Normative reasons, we have seen, seem to be worldly facts concerning the actual desirability of actions. According to the perspectivalist view, the explanatory reasons that make sense of actions from the agent's point of view are fundamentally separate from these normative considerations. According to Johannes Roessler (2014), this 'sever[s] or at least complicate[s] the link' between practical questions about what to do and how, and questions about why one does what one does. As we saw in the previous chapter, such questions seem from the agent's perspective to come together: one reasons what to do on the basis of reasons that one then takes to explain one's doing what one does. On a perspectivalist view, Roessler argues, one 'would have to think about one's action from a standpoint that is neutral on whether one is getting things right about one's practical reasoning' (Roessler, 2014, p. 351). This seems false to the way we typically do think about our own actions, and Roessler appears to think that we must endorse factivism in order to avoid this unattractive conclusion.

It is not clear, though, why a perspectivalist view would imply that we would have to think about our own actions in this way. If we take the view of perspectival rationalisation I outlined in the previous chapter, it seems that from the agent's perspective the two questions Roessler identifies do go together. What the perspectivalist view insists upon is just that it is part of our understanding of rational agency that agents can make mistakes of a sort that do not make a difference to the rational intelligibility of their actions. This does not mean that when we deliberate we do so from a standpoint that is neutral on whether we are getting things right. After all, the agent's aim in deliberating is not merely to act in a way that is rationally intelligible, but to act in a way that she actually has good reason to act—and as we have seen, nothing about the perspectivalist view commits one to denying that what one has good reason to do is a question that must be answered by considering worldly reasons.

[51]. Or, as in Martin's version of object-perception disjunctivism, that the common kind is fundamentally to be understood negatively in relation to the kind special to the 'good' case, so that the common kind, while common, lacks explanatory autonomy.

2.4.2 Williamson's non-conjunctivism about knowledge

A number of authors have argued that in order for a worldly rationalisation to be available, the agent must know the fact that is to rationalise their action.[52] If this is so, perhaps we might make a case for the distinctiveness and autonomy of worldly rationalisation on the basis of the distinctiveness and autonomy of knowledge with respect to belief. The idea that knowledge is autonomous has recently gained a great deal of traction in epistemology. 'Knowledge-first' theorists reject the traditional project of seeking to define knowledge in terms of belief, and take knowledge to be a quite distinctive state with an explanatory role not reducible to that of belief.[53] The factivist says that acting for a reason is not to be understood in terms of treating-as-a-reason. If the former requires knowledge, then perhaps we might be able to motivate a factivist view on the basis of the knowledge-first thesis that knowledge cannot be understood in terms of belief. Worldly rationalisation would be shown to have a kind of explanatory autonomy derivative of the explanatory autonomy of knowledge. However, I think the prospects of arriving at an interesting factivist position by this route are actually rather dim.

Perhaps the most obvious way to expand the knowledge-first enterprise in epistemology to encompass a factivist approach to rational action would be to say something like the following. For someone to *V* because they think that *p* requires that they think that *p*. And as John Hyman, Jennifer Hornsby and others argue, for someone to *V* because *p* requires that they know that *p*. Given that the fact cited in explaining the agent's action in the former case is a fact about the agent's believing something, it is not unnatural to think that we should understand perspectival rationalisation in terms of belief. Similarly, one might argue, what is really explaining the agent's *p*-ing when we give a worldly rationalisation is that the agent knows that *p*. As perspectival rationalisation is essentially understood in terms of the psychological operation of the state of belief, worldly rationalisation is essentially understood in terms of the psychological operation of the state of knowledge. If this is right, and if knowledge is autonomous as the knowledge-first epistemologist claims, then worldly rationalisation will be, derivatively, autonomous too.

The major challenge for this approach will be to explain how we should understand the two forms of rationalisation in terms of knowledge and belief. It is not at all clear that we can say what it is to act for a reason, or what it is to act for an apparent reason, in terms of knowledge and belief respectively. A natural way to do so would be to posit some distinctive kind of 'rational causation', and to say that (for the worldly case) someone *V*-s because *p* just in case their *V*-ing is caused in the relevant way by their knowing that *p*. It will need to be explained what the relevant kind of causation is—if it is left primitive, then we are effectively leaving the notion of acting for a reason primitive, and so not really explaining acting for a reason in terms of knowledge at all. The prospects for this project, though, look very poor. Notoriously, theories of this sort run into the problem of 'deviant causal chains', which I

[52]. (Hornsby, 2008; Hyman, 1999; McDowell, 2013; Unger, 1975; Williamson, 2000)
[53]. See (Williamson, 2000 for the canonical statement of the knowledge-first approach).

discussed briefly at the end of the last chapter. These are counterexample cases in which the specified conditions for the 'right kind of causation' are met, but in which, intuitively, the agent does not act for the relevant reason.[54] Interestingly, the deviant causal chain problem is remarkably similar to the Gettier problem in epistemology, and the failure of the traditional epistemological project to satisfactorily answer the latter problem is a major part of the motivation for knowledge-first epistemology. If the intractability of the Gettier problem justifies treating knowledge as primitive, it is hard to see why the problem of deviant causal chains should not equally justify taking the notion of acting for a reason as primitive rather than trying to analyse it in terms of a specific kind of causal connection between mental state and action.

However, if our method of expanding knowledge-first epistemology into a factivist theory of rational action is not reductive in something like the way considered above, then it is not clear that it will have anything distinctive to contribute to our enquiry concerning the two forms of rationalisation and the connection between them. If we look for an account based on weaker claims than the reductive ones just considered, there are a couple of aspects of the connections between worldly rationalisation and knowledge, and between perspectival rationalisation and belief, that we might focus on. One is the idea that an agent's being in the relevant state is merely a necessary condition for the availability of the relevant rationalisation. Pairing this with a knowledge-first epistemology would not in itself constitute a factivist theory of rationalisation. A perspectivalist theorist could in principle hold that a worldly rationalisation is available just in case an agent does something on the basis of their belief that *p*, whilst knowing that *p*. How informative such a theory would be would depend, of course, on whether we could give some kind of substantive positive account of doing something on the basis of a belief, but that we could is by no means ruled out by the claim that knowledge is autonomous with respect to belief.

The other thing that the knowledge-first theorist might want to say is that one's *V*-ing because *p* just is one's *V*-ing's being caused in the right way by one's knowledge that *p*, but that the 'right kind' of causation is not something of which we could give an independent or reductive account.[55] This would be a factivist view since it would entail that we could not give an independent or reductive account of worldly rationalisation in terms of perspectival rationalisation. However, it is not clear what the account gains from its token association with knowledge-first epistemology. In order to have a reason to accept the account, we need some argument as to why we should not think of 'the right kind of causation' in the knowledge case as being explicable in terms of a kind of causation common between the cases. In other words, we are still without an argument for factivism.

[54]. See (Davidson, 1980b, Chapter 4). For a nice overview of the ways in which the deviant causation problem arises for different accounts of rational causation, see (Mayr, 2011, Chapter 5).

[55]. This would be akin to Davidson's treatment of deviant causal chains, but within a knowledge-first, factivist framework.

2.4.3 Hyman's account of knowledge and belief

One place where I think a distinctively factivist view might be found is in the work of John Hyman. Hyman argues not just that knowledge is a necessary condition on someone's responding to a worldly reason, but that knowledge should be understood essentially in such terms. For Hyman, knowing that *p* simply is having the ability to be guided by the fact that *p*, where action (belief, judgement, …) that is guided by the fact that *p* is action (belief, judgement, …) that can be given a worldly rationalisation in terms of the fact that *p*.[56] So for Hyman, knowledge itself is to be understood in terms of the kind of relation imputed by a worldly rationalisation.

This view is in itself compatible with a perspectivalist view. It could be that knowledge is to be understood in terms of the idea of being guided by a fact, which is itself understood in terms of acting on the basis of a belief or for an apparent reason. What makes Hyman's account factivist is its combination of the above account of knowledge with a distinctive account of belief: believing that *p*, Hyman suggests, is being disposed 'to act (think, feel) as one would if one knew that *p*, or as one would if one were guided by the fact that *p*' (Hyman, 2015, p. 173). Hyman's theory thus gets at perspectival rationalisation from worldly rationalisation, via knowledge and belief. It is factivist in that it holds perspectival rationalisation to be explicable in terms of worldly rationalisation: knowledge is defined in terms of doing things for reasons, belief is defined in terms of knowledge. Presumably perspectival rationalisation is explained in terms of belief: one's *V*-ing is explicable by a perspectival rationalisation when it manifests one's disposition to behave as one would if one knew.

To assess Hyman's account, we need to assess both his claim about knowledge and his claim about belief. The account of knowledge as the ability to be guided by the facts depends in part on supporting the weaker claim that an agent can be guided by the fact that *p* (so that '*p*' can be given as a worldly rationalisation of her action) only if she knows that *p*. Hyman (1999) and Hornsby (2008) argue for this claim in much the same way. The argument takes the form of an inference to the best explanation. We can begin with some basic observations. For someone to act because *p*, it must seem to them that *p*, in much the same way as it must for them to act because they think that *p*. Moreover, it must be the case that *p*, since, in general, '*q* because *p*' entails both that *q* and that *p*. However, the conjunction of these conditions is not enough to make possible a worldly rationalisation of someone's acting in terms of the fact that *p*. The mere truth of what one believes does not guarantee that there is any explanatory connection between that truth and one's action. A potential fix might be to say that one's belief must be justified, or based on good reasons. As Hyman and Hornsby observe, though, such an account fails for worldly rationalisation just as it fails for knowledge, as is shown by Gettier cases. Hornsby gives the following example:

> Edmund … believes that the ice in the middle of the pond is dangerously thin, having been told so by a normally reliable friend, and … accordingly keeps to the edge. But Edmund's friend didn't want Edmund to skate in the

[56]. The view is developed in (Hyman, 1999, 2006, 2010, 2011, 2015).

> middle of the pond (never mind why), so that he had told Edmund that the ice there was thin despite having no view about whether or not it actually was thin. Edmund, then, did not keep to the edge because the ice in the middle was thin. Suppose now that, as it happened, the ice in the middle of the pond was thin. This makes no difference. Edmund still didn't keep to the edge because the ice was thin. The fact that the ice was thin does not explain Edmund's acting, even though Edmund did believe that it was thin, and even though the fact that it was thin actually was a reason for him to stay at the edge. (Hornsby, 2008, p. 251)

The unavailability of worldly rationalisation in such cases calls for explanation, and Hyman and Hornsby suggest that the best explanation is that, for one to *V* because *p* (in the sense of worldly rationalisation), one must know that *p*.

This argument has recently been challenged. Nick Hughes (2014) and Dustin Locke (2015) (independently) appeal to cases structurally similar to the well-known 'fake barns' case,[57] and argue that in such cases an agent can act for a reason he does not know. Here is Hughes:

> Henry is out hiking. He's lost, and the weather is turning nasty. The situation is getting serious. He sees what he believes to be a hiker's hut in the distance, and feels relieved. In fact, unbeknownst to Henry, he is in fake hiker's-hut county—an area where there are only a handful of real huts, and many hut-facades designed to look exactly like real huts to passing hikers. Henry justifiably and truly believes that the structure in the distance is a hut, but he does not know this. (Hughes, 2014, p. 461)

Hughes suggests that in this case, 'Henry feels relieved because there is a hut in the distance' would be a legitimate, and genuinely rationalising, explanation of Henry's feeling relieved. If that's correct, then while knowing that *p* might well be sufficient to put one into a position to do things because *p*, it would seem that it is not, in general, necessary.

The claim that worldly rationalisations are available in fake-barn-type cases is controversial.[58] With our focus on the relationship between worldly and perspectival rationalisation, though, we can for now sidestep this issue. What is required for an agent to be in the position to respond to the fact that *p* is that the agent be *aware* that *p*, where awareness that *p* is a factive cognitive condition. It may be that being aware of a fact just is, after all, knowing that fact. Alternatively, it might be that knowledge is more demanding than awareness: perhaps knowing that *p* is a matter of both being aware that *p* and believing that *p* with the right kind of normative warrant.[59] For now, we are simply trying to see whether a plausible factivist account can be developed from Hyman's ideas. We can investigate that question by simply substituting 'awareness' for 'knowledge' in his account, remaining neutral on whether awareness in the relevant sense just is knowledge.

The factivist character of Hyman's account depends just as much on the account of belief as it does on the account of knowledge. Hyman does not exactly argue for his account of belief so much as present it as a plausible suggestion. Here I think we get into more trouble, because it is not clear that we can understand believing that *p* in terms of being

[57]. The case originates in (A. I. Goldman, 1976).
[58]. See (Cunningham, 2018; Littlejohn, 2014) for responses.
[59]. Compare the account of memory-based knowledge in (Peacocke, 1986).

disposed to act as one would if one was aware that *p*. A straightforward kind of challenge to this analysis exploits precisely the feature of awareness that we are interested in, namely its factivity. Consider the following pair of cases. Krzysztof is a world record-holding powerlifter. He is aware that he can deadlift 420kg. Krzysztof is in a powerlifting competition in which he needs to deadlift 375kg to win. Aware that he can deadlift considerably more than this, Krzysztof is disposed to (successfully) deadlift the 375kg, and to do so with ease. Christopher is not a world record-holding powerlifter. In fact, he is not much of a powerlifter at all. He is, however, delusional, and he believes that he can deadlift 420kg. Is he disposed to act as he would if he were aware of the fact that he could deadlift 420kg? In some respects, yes. Perhaps he is disposed to enter powerlifting competitions and to feel very confident about his chances, for example. But he lacks other dispositions that come with the kind of awareness that Krzysztof has. Notably, he is not disposed to actually, successfully, deadlift 420kg, or 375kg, or even 180kg. Because one cannot be aware that *p* without its being the case that *p*, and because often its being the case that *p* will make a difference to what one can do or is disposed to do,[60] it seems that we cannot say that in general belief disposes one to act as one would if one was aware.

Perhaps Hyman's account of belief can be amended so as to avoid this kind of challenge, but perhaps not. As it stands, we have seen no positive reason to think that we have to understand belief in terms of knowledge, and we have seen that there is trouble with Hyman's specific attempt to do so. Hyman's account does offer a model both of what a factivist view might look like and of how it might be motivated (namely by making a compelling case for the accounts of knowledge and belief in which it consists). I will not attempt to develop Hyman's view in this way, though. I believe a more direct case can be made for the factivist view, as we can see if we return to consider the McDowellian epistemological project discussed earlier.

2.5 Factivist epistemology

Factivism is the rejection of the perspectivalist view. The perspectivalist theorist holds that while worldly rationalisation, in which we explain a person's action (belief, judgement, feeling, ...) by straight statement of a fact that was a reason for them to act as they did, is an explanatory form that is only available when the person was aware of the stated fact, the factivity of the explanation makes no significant difference to the manner of our understanding of what the person did. In rejecting this claim, the factivist holds that, at least in some cases, an action's (belief's, judgement's, feeling's, ...) being directly explained by a fact about how things stand, a fact that was a good reason for them to do what they did, makes sense of what they did in a way that cannot be captured by explanations that do not entail that there was such a reason. The most direct kind of argument for factivism will be one that supports the claim that, in a specific kind of case, the worldly rationalisation must be treated

[60]. (Gibbons, 2001 argues for an even tighter connection between knowledge and abilities to act than the one illustrated by my example cases, namely that one cannot intentionally V without knowing how to V).

as basic. One way to support this will be to show that the way in which the corresponding perspectival rationalisation makes sense of the action (belief, ...) has to be understood in relation to the worldly rationalisation. We have already noted a connection between the factivist view concerning rationalising explanation and McDowellian disjunctivism about perceptual knowledge. In this section, I will argue that a factivist position can be motivated precisely as an attractive development of that form of disjunctivism.[61] The case for factivism thus made will therefore be only as persuasive as the disjunctivist approach to perceptual knowledge itself. Given the close connection between the two views, though, this is perhaps to be expected.

I want to think about the justification that one can have for believing that *p* in virtue of its visually seeming to one that *p*. For simplicity, I will refer to a belief that is justified in this way a 'perceptual belief'. I take it that most will want to agree that many of our beliefs are justified in this way. Where we tend to find more disagreement is when we ask how exactly appearances justify beliefs.

It is natural to suppose that whenever someone who believes that *p* is justified in believing that *p*, she has a justification for believing that *p*, and this justification can be represented abstractly in the form of an argument to the conclusion that *p*. Moreover, for this justification to be what actually justifies the agent's belief, it must somehow correspond to the actual cognitive basis of the agent's having that belief, in a way that is reflected in a correct rationalisation of the belief. For example, Eugène might justifiably believe that Isidore is mortal on the basis that

> Isidore is a cat
> All cats are mortal
>
> ———————————
>
> Isidore is mortal

where Eugène believes both premises and believes the conclusion because he believes the premises.[62]

There is a puzzle about how to apply this principle to the justification of perceptual belief. That is: When it looks to me as if *p*, and I justifiably believe that *p* as a result, what is the justification of my believing that *p*? The justification must be something provided by my experience: but what premise or premises, provided by my experience, might justify my believing that *p*? One obvious candidate for a first premise is that it looks to me as if *p*. Patently, this cannot be the whole of my justification, since its looking to me as if *p* is consistent with its not being the case that *p*. So the premise that it looks to me as if *p* has to be supplemented with some further premise or premises that somehow bridge the gap between appearance and reality, delivering the conclusion that *p*. Suppose I believe that there is a barn over there because it looks to me as if there's a barn over there. Its looking to me as if

[61]. Something along these lines is suggested by (Hornsby, 2008), but the argument is not developed in detail.

[62]. Compare (Harman, 1964).

there is a barn over there is quite consistent with there being no such barn. What looks to me like a barn might be a mere barn-façade, or I might be hallucinating a barn. The gap between appearance and reality is a gap in our justifying argument:

It looks as if there is a barn
…
_____
There is a barn

What might fill the gap?

One possibility is that the argument is abductive. So perhaps the argument is

It looks as if there is a barn
The best explanation for its looking as if there is a barn is that there is a barn
_____
There is a barn

Or perhaps it is inductive:

It looks as if there is a barn
Normally when it looks as if there is a barn, there is a barn
_____
There is a barn

Here we face some trouble, however. If the second premise is to justify my believing that there is a barn, it must not only play the right kind of role in grounding that belief, it must also play that role justifiably. This seems to require that insofar as my perceptual belief that there is a barn is based on premises that I believe, I must be justified in believing the premises if my perceptual belief is to constitute knowledge. Presumably I am justified in believing the first premise because I am conscious of how things appear to me, but what is my justification for believing that there being a barn is the best explanation of its looking to me as if there is a barn, or that reality usually comports to my experiences?

It seems that to be justified in believing any premise that would bridge the gap between appearance and reality, I would need already to be justified in having certain kinds of general beliefs about how things work in the external world, how reliable my senses are, and so on. Now, assuming that I have such a justification, one or the other of the above arguments might be capable of justifying my belief that there is a barn. The trouble is that we are looking for a general account of how perceptual experience provides us with knowledge of the external world, and it is very hard to see how one could be justified in believing a general proposition about how the world works of the kind featuring in our inductive and abductive arguments above, except on the basis of experience. To know that its being the case that *p* is

usually the best explanation for its appearing as if *p*, or that when it appears as if *p* it normally is the case that *p*, one needs to know something about how appearances tend to connect with reality, which one can only know if one knows about how things tend to work in reality, and it is deeply obscure how one could come by that kind of knowledge without already being capable of knowing things on the basis of perceptual experience.[63] So it is very hard to see how we could get from premises concerning only how things appear to conclusions about how things really are if this were the only way perceptual experience could provide justification for perceptual beliefs.

Some authors, faced with this puzzle, favour rejecting the principle about justification. Perhaps perceptual beliefs are *immediately* justified, where A is immediately justified in believing that *p* just in case A is justified in believing that *p*, but is not so justified in virtue of 'some relation this belief has to some other justified belief(s)' of A's (Alston, 1983, p. 74). Read strictly, I think that this is in fact the correct response to the worry: my belief that there is a barn over is not justified by inference from any other beliefs. However, proponents of 'immediate justification' typically mean not just that perceptual beliefs are not rationally grounded in other prior beliefs, but that perceptual beliefs do not need a justification in the sense outlined above at all.

Jim Pryor, for instance, argues that having a justification for believing does not mean that you must 'always be able to offer *reasons* ... in support of your belief' (Pryor, 2000, p. 535). Similarly, Clayton Littlejohn rejects the principle because it seems false as applied to action: it can be reasonable to perform actions that we have no reason at all to do, such as idly doodling (Littlejohn, 2015). However, while both observations seem plausible, neither succeeds as a motivation for rejecting the principle about justification. While Pryor is right that we need not always be able to offer reasons in support of our beliefs, to think that this speaks against the principle is just to confuse having a justification in the sense I described above—one's belief's having a cognitive basis that corresponds to a good argument for the beliefs' content—with something quite different, namely being able to justify one's belief to someone else. The latter appears much more demanding than anything that our principle about justification commits us to. The importance of this distinction will come out more clearly shortly.

Littlejohn's argument fails for a different reason: he assumes that there is a relevant analogy here between belief and action, whereas there is in fact good reason to expect a disanalogy. There are plausible explanations as to why activities like idle doodling do not typically require justification. They are harmless and virtually cost-free. There is no particular reason *not* to engage in them, assuming one does not have much stronger reason to be doing something else (in which case doodling might well require justification). We tend to think differently about belief: there seems to be a presumption against believing things that are not true,[64] such that justifiably believing something requires one to have some reason to think that it is not false. If there were some general norm of action to the effect that we ought never

[63]. See (McDowell, 1994, 1995, 1998a).

[64]. For the idea that 'truth is the norm of belief', see for instance (Engel, 2013; Littlejohn, 2012; Shah, 2003; Wedgwood, 2002).

to do anything that we do not have positive reason to do, akin to the totalitarian principle that everything is forbidden unless expressly allowed, then perhaps activities like idle doodling would always need a justification. Applied generally to action, such a principle seems absurd. The reason it is perfectly alright to doodle for no reason is that one needs no justification for doodling. This does not show that idle doodling is justified despite not being done for a reason. Moreover, we do not ask something like doodling to provide the justificatory basis for whole systems of ends, whereas we do ask perceptual beliefs to provide the justificatory basis for whole systems of belief. In light of this it should hardly be surprising if the latter are held to a normative standard that the former are not.

2.5.1 Believing in light of a fact

We have seen that there are serious problems both with saying that perceptual beliefs have no justification and with saying that the justification they do have is based on premises about how things appear to the subject. In light of this, we should consider whether a better premise might be available. Recall the contrast, discussed in the previous chapter, between the way that the fact that one believes that *p* rationalises action when that fact is itself one's reason for acting, and the way the fact that one believes such-and-such rationalises action when it does so as a perspectival rationalisation. The same kind of contrast applies to the rationalisation of perceptual belief in terms of facts about how things appear to a perceiving subject. Just as it would be a mistake to see ordinary perspectival rationalisation in the case of action as giving the agent's reason for acting, I think it is exactly the same kind of mistake to see a typical rationalisation of the form 'A believes that *p* because it looks to her as if *p*' as giving, in the clause following the 'because', A's reason for believing that *p*. Just as in the case of action A's action is justified, from her point of view, by *what* she believes, I want to suggest that what justifies A's perceptual belief from her point of view is not its appearing to her as if *p*, but simply: that *p*. On this account, the canonical justifying argument for a perceptual belief that *p* is the simplest kind of argument there is. It just relies on the repetition rule. The argument is: *p*, therefore *p*.

If this is right, then in the good case, where I genuinely see that *p* and I believe that *p* on the basis of my experience, the canonical worldly rationalisation of my so believing is: 'He believes that *p* because *p*'. The perceptual experience thus provides a premise which, when true, is a conclusive reason for the perceptual belief. There is no logical gap to be bridged between appearance and reality, because the premise of the argument itself concerns reality.[65]

I will argue that only a factivist view can make sense of the structure of justification here. First, though, I want to try to allay some potential worries about the suggestion. An immediate worry might be that the justificatory argument '*p*, therefore *p*' is question-

[65]. While this proposal is McDowellian in spirit, it differs from McDowell's own account, which holds that one's reason for believing that *p* is the fact that one sees that *p*. I think the version given here is preferable in that it does not take all knowledge of the external world to be inferred from reflective self-knowledge of something logically stronger. It also means that we do not need to give an account of how the perceiving subject knows that she sees that *p* without already knowing that *p*. And, I think, it comports better with the central idea of McDowell's account: that in veridical perception, a worldly fact is made manifest.

begging. Such an argument certainly might be question-begging if offered as an attempt to convince someone else that *p*, but as I noted above, the ability to convince others is not what we are concerned with here. Neither are we concerned to refute scepticism. What we are seeking is an explanation of how perceptual beliefs could have a cognitive basis that gives them the normative status necessary for them to constitute knowledge. For these purposes, the 'question-begging' nature of the argument is not problematic, indeed it is its primary virtue. There could hardly be a better justification for believing that *p* than the very fact that *p*.

A related worry is that the *explanation* 'She believes that *p* because *p*' is circular, or not a real explanation. This would be true if a crude version of the perspectivalist view, on which every worldly rationalisation is to be understood in terms of a belief-ascribing perspectival rationalisation, were true. If it were, then the rationalising content of 'She believes that *p* because *p*' would be given by 'She believes that *p* because she believes that *p*', which is plainly untenable. To hold that perceptual beliefs are justified by an argument that employs the repetition rule, then, we must reject the idea that the cognitive basis of a belief—the states or attitudes on the basis of which one believes, whose contents correspond to the premises of the justifying argument—must always consist solely in beliefs. So the subject's perceptual awareness that *p* cannot be understood in terms of her believing that *p*. Littlejohn rejects the repetition view on just these grounds, arguing that being aware that *p* involves the exercise of conceptual capacities, an exercise which is 'distinctive of belief' (Littlejohn, 2015). Again, the argument is unconvincing. Conceptual capacities are not only exercised in belief: they are also exercised in supposition and linguistic comprehension, for example. It is unclear why we should not take perceptual awareness of a fact to constitute another exercise of conceptual capacities that is not itself belief.[66] Moreover, if being perceptually aware that *p* involves the exercise of conceptual capacities, so too, presumably, does having it perceptually appear to one that *p*. Plainly, one can be in a situation wherein it perceptually appears to one that *p* without one's believing that *p*: this is just the situation one is in when one is knowingly subject to an illusion or a hallucination. So it is quite plausible that having a perceptual appearance as of its being the case that *p* can be causally and rationally antecedent to believing that *p*. If it appears to one that *p* because it is the case that *p* (which, in the good case, it does), there seems to be no obstacle to saying that one can, in that case, believe that *p* because *p*.

The suggestion that perceptual beliefs are justified by application of the repetition rule is plainly inconsistent, then, with a crude perspectivalist view on which the only perspectival rationalisations are belief-ascribing ones. The cognitive basis, the state or attitude that causes the perceptual belief ('in the right way') must be the perceptual experience itself: the state of having it perceptually appear to one that *p*. As we might naturally put it, forming the perceptual belief is not a matter of inferring its content from other things one believes, but of

[66]. This is, I think, consistent with denying that the phenomenal character of perceptual experience is fundamentally to be understood in terms of conceptual content. Even if we endorse a 'nonconceptual content' view or 'object view' of perception, we will still need to make sense of the idea of seeing-that, and seeing-that seems to involve the application of concepts.

simply accepting the appearance. The perspectivalist theorist, then, can try to endorse the present suggestion about the justification of perceptual belief by saying that the worldly rationalisation 'She believes that *p* because *p*' is to be fundamentally understood in terms of the perspectival rationalisation 'She believes that *p* because it appears to her that *p*'.

There is nonetheless reason to doubt whether even this more sophisticated perspectivalist view can accommodate our suggestion about how perceptual beliefs are justified. On the perspectivalist view, the rationalising import of 'A believes that *p* because *p*' must be understood in terms of the rationalising import of 'A believes that *p* because it appears to her as if *p*'. However, if we consider the phenomenon of defeat, it appears that this order of explanation needs to be reversed.

Suppose that Henry is in fake hut country. It looks to him as if there is a hut over there on the hillside, and, in virtue of this perceptual appearance, he is justified in believing that there is a hut. Now suppose that Henry finds out that he is in fake hut country, perhaps because a friend tells him that he is. Henry loses his justification for believing that there is a hut over there on the hillside. Why?

We have here a pair of cases. We might, considering them as it were from Henry's perspective, call them the *apparent good case* and the *possible bad case*. In both cases, Henry sees an object O, which looks like a hut. In the apparent good case, the 'premise' of Henry's justification for believing that O is a hut is: that is a hut.[67] In the apparent good case, Henry's belief that O is a hut is justified; in the possible bad case it is not. The thing that makes a difference to Henry's epistemic situation between the two cases is that in the latter, Henry knows that he is in fake hut country. That knowledge, it seems, defeats whatever justification Henry had for believing that O is a hut.

How does the knowledge that he is in fake hut country defeat his justification for believing that O is a hut? If we think, again, of a justification as representable in argument form, we can see three ways in which a justification might be defeated. First, a defeater might be a reason to believe the negation of the conclusion. Second, it might be a reason to doubt one or more of the argument's premises. Third, it might call into question the connection between the premises and the conclusion.[68]

The fact that Henry is in fake hut country does not defeat his justification in the first way. Not everything that looks like a hut in fake hut country is a fake hut. There are real huts too. While the fact that one is in fake hut country provides some statistical evidence that any given hut-looking thing one sees might not be a hut, Henry's justification for believing that O is a hut was not statistical or inductive. Henry did not infer from the fact that O looked like a hut to the conclusion that it is a hut. His justification for believing that O is a hut was: that is a hut. If this justification is true, it is a conclusive reason, and cannot be rebutted by merely statistical evidence. If you can see that the swan before you is black, your justification for believing as much is not undermined by the fact that most swans are white and that this,

[67]. Note that in speaking of a 'premise' here I do not mean to say that Henry's belief is based on an *inference*, if this is thought to involve basing a belief on other things one believes, or going through some kind of conscious reasoning.

[68]. Compare the distinction between rebutting and undercutting defeaters in (Pollock, 1986).

being a swan, is therefore probably white. The fact that most swans are white is simply irrelevant.

The fact that Henry is in fake hut country also cannot defeat his justification in the third way, by calling into question the connection between the premise and the conclusion. Not only is the connection between premise and conclusion one of entailment, it is the clearest and most straightforward kind of entailment possible: a proposition's entailing itself.

So it seems that Henry's being in fake hut country must defeat his justification by somehow 'attacking' its premise, the premise that that (O) is a hut. Where the cognitive basis of a belief consists in other beliefs, one way of defeating the conclusion-belief is by undermining whatever rational support or justification one has for one or more of the premise-beliefs. If Eugène believes that Isidore is mortal because he believes that Isidore is a cat and that all cats are mortal, for instance, we might defeat his justification for this by providing him with strong evidence of the existence of immortal cats. This would suggest that the second of his premise-beliefs is not altogether kosher, and that he is not entitled to believe anything else on the basis of it. This cannot be exactly what is going on in Henry's case, however, because the cognitive basis of his belief that O is a hut—the premise-state, as we might call it—is not a belief but a perceptual appearance. Unlike beliefs, perceptual appearances are not based on any rational grounds; they are not justified and do not require justification. One is simply subject to them, and when one is subject to them, they are in principle apt to justify beliefs. So Henry's justification cannot be defeated by something's showing his premise-state to be unjustified.

The only explanation remaining seems to be that the fact that Henry is in fake hut country is a reason to doubt whether his premise-state—the state of its visually appearing to him that O is a hut—is the right kind of state to provide adequate grounds for knowledge. The perspectivalist theorist, it seems, has no explanation available to them of why this should be. For while the perspectivalist theorist can make a distinction between seeing that *p* and merely seeming to see that *p* (that is, having a visual appearance as of its being the case that *p* but not actually seeing that *p*), the difference between the two states does not, on the perspectivalist view, make a difference to the way in which the subject's seeming to see that *p* (that is, the condition common between good and bad cases) provides the subject with rational grounds. The perspectivalist thesis is precisely that we should understand the way worldly reasons rationalise in terms of the way that non-factive mental states rationalise. Regarding the possible bad case, we want to say that the non-factive mental state of having things visually appear that *p* cannot adequately rationalise one's believing that *p*. The only explanation for this, though, seems to be that in the possible bad case, one is aware that one might not be in a position to believe that *p* because *p*—that were one to believe, one's so believing might not be rationalised by how things really are. This also suggests that in the actual bad case, where one believes that *p* merely because it seems to one that *p*, one's justification for believing that *p* depends on the fact that it is for one as if one is believing that *p* because *p*.

The perspectivalist thus cannot make sense of Henry's own perspective on his situation. What he learns, when he learns that he is in fake country, is that his belief that that is a hut might not be rationalised in the way that he had, in basing that belief on how things looked to him, implicitly assumed that it was. He learns that he might not be believing that that is a hut because it is a hut. This suggests that we need to understand the rationalising significance of the perspectival rationalisation 'he believes that *p* because it looks to him as if *p*' in relation to the worldly rationalisation 'he believes that *p* because *p*'. The former only rationalises because it tells us that it is for the subject as if the latter is true.

2.5.2 Generalising the argument for factivism

I have argued that the justificatory structure of perceptual appearances is best accounted for if we take the case in which someone believes that *p* because *p*, where the fact that *p* is their reason for believing that *p*, to be fundamental, and understand the rationalising role of the mere appearance that *p* in relation to that primary case. As it stands, this argument only makes a case for the fundamentality of worldly rationalisation as applied to perceptual beliefs. It is worth considering whether and to what extent the conclusion might be generalised: first, whether its application to the rational grounding of theoretical knowledge might be broadened; second, whether there is a case for thinking that there is ever a similar relation between worldly and perspectival rationalisations in the rationalisation of action.

While the argument for factivism given above specifically discusses perceptual knowledge and the justification of perceptual belief, the aspects of perception upon which the argument depended are, plausibly, not unique to perception. The key features of perception were: first, that the best way to make sense of the justification provided by perception involved seeing the 'premise' of the justificatory 'argument' as the very proposition believed, so that in the good case the worldly rationalisation of the subject's belief that *p* is: she believes that *p* because *p*; second, that this justification is subject to a certain kind of defeat when the subject has reason to believe that she might not be in the good case.

Insofar as these features do in fact characterise perceptual knowledge, it seems plausible that they will also characterise other kinds of knowledge. Specifically, wherever a subject's knowledge is based on the actualisation of a basic epistemic capacity that delivers a 'seeming' as of something's being the case, we might expect the same kind of structure to be present. In these cases, the same kinds of sceptical worries about the nature of the subject's justification are liable to arise, so that we should be able to make the same kind of argument for treating the justification for her belief as employing the repetition rule. And in these cases we can expect to be able to generate examples where the subject's justification is defeated in the same type of way as in the 'fake barns' case considered above. More precisely, it seems plausible that the structure will apply wherever there is a kind of basis for knowledge that is both foundational and fallible. This might include not just perception but also memory, introspection, 'intuition' of basic principles of reasoning, and perhaps even testimony. Of

course it might be that issues arise in each or any of these cases that make it more difficult to apply the factivist approach developed above, and developing a broad factivist account of knowledge would take much more work than is feasible here. The point I want to make is just that there is at least some reason to expect that the features of perceptual knowledge that made the argument for factivism possible are not unique to sensory perception.

This observation also raises prospects for developing a distinctively factivist view about the rationalisation of action. If we assume that there are facts of the matter about which facts constitute reasons for which kinds of action, questions arise about how we know what reasons we have and how we manage to respond appropriately to those reasons. In seeking to answer such questions, we might well be attracted to an account according to which our competences to respond to reasons consist or are grounded in some kind of basic sensitivity to reasons, akin to basic epistemic capacities like perception. On such an account, the starting-point for practical reasoning might be a state in which some consideration seems to the agent to be a reason to act in a certain way.

One author who accords an important role to the idea of something's seeming to one to be a reason is T. M. Scanlon, in particular in *What We Owe to Each Other* (Scanlon, 1998). Although it is not entirely clear from what Scanlon says whether he thinks we should see the rationalising role of seeing-as-a-reason as being relevantly analogous to that of (literally) seeing that something is the case, he does emphasise that these seemings arise independently of one's judgements and are often recalcitrant to those judgements, and he says that '[s]eeming to be a reason is … a matter of appearing to be one' (Scanlon, 1998, p. 65). We might well, then, read Scanlon as proposing that these practical seemings play a role in rationalising action at least somewhat analogous to that of perception in rationalising perceptual belief.[69] So there might, perhaps, be material for a distinctively practical factivism in a Scanlonian view.

For Scanlon, seeming to have a reason to *V* is central to being rationally motivated to *V*. Indeed, to have what is ordinarily called a *desire* to *V* is, on Scanlon's account, to have some consideration or set of considerations 'insistently' appear to one as (a) reason(s) to *V*. His account therefore promises to account for another central form of rationalisation—the explanation of action in terms of a desire of the agent's—in terms of worldly or perspectival rationalisation. 'A *V*'d because she had a desire to *V*' rationalises, on a Scanlonian account, by telling us that, as it seemed to A, she had a worldly reason to *V*. Whether or not the argument for factivism were to go through, this would make for an attractively unified conception of rationalisation: either we would understand both perspectival and desire-ascribing rationalisations in terms of worldly rationalisations, or we would understand both worldly and desire-ascribing rationalisations in terms of perspectival rationalisations. Either way, all reasons for action could be understood as worldly and universal. In the next chapter, however, we will see that there are serious challenges for a Scanlonian conception of desire. These challenges, I will argue, suggest that we do not in fact understand our own reasons for acting in purely worldly and universal terms.

[69]. See also (Stampe, 1987), who makes this connection more explicitly.

Chapter 3

The Wings of Desire

3.1 The idiosyncrasy of desire

I propose, in order to investigate the question of the rationalising role of desire-ascription effectively, to set aside the factivist–perspectivalist distinction. We saw in Chapter 1 how worldly and perspectival rationalisations can both be understood in terms of universal worldly reasons, in a way that is in principle neutral on the order of conceptual priority between the two explanatory forms. We can now frame another question, taking our lead from the Scanlonian picture introduced at the end of Chapter 2: is the rationalising role of desire to be understood in terms of worldly reasons?

Like perspectival rationalisation, rationalisations of action in terms of the agent's desires introduce a kind of idiosyncrasy. Different people desire different things, and it seems that what a person desires often plays an important role in explaining that person's actions: sometimes a person's desires make sense of actions that the person's worldly reasons cannot. On what I will call the *cognitivist* view, the nature of the idiosyncrasy introduced by desire is essentially the same as that of the idiosyncrasy of belief: it is an idiosyncrasy of the agent's perspective on what worldly reasons they have, the idiosyncrasy of apparent reasons, considerations that seem to the agent to be reasons. Cognitivism is consistent with both perspectivalist and factivist views about worldly and perspectival rationalisation. On the cognitivist view, we still understand our reasons as being essentially worldly and universal. Desire is not a special, non-worldly source of reasons. It is just that in desiring it can seem to us that we have reasons that we do not really have.

In this chapter, I will argue that cognitivism struggles to accommodate a certain kind of rational or at least not-irrational motivation. This kind of motivation is not a marginal phenomenon, but holds a central place in the good life for most if not all people. It seems to constitute an important source of reasons that are importantly different in their rationalising character from worldly reasons. These reasons are personal and particular rather than worldly and universal. Or so I will suggest.

3.1.1 Scanlon's cognitivist model

Since Scanlon will be our model cognitivist in what follows, it will be helpful to have the basics of his picture of reasons and rational motivation clearly in view.[70] A few features of the view will be particularly important for our discussion. First, reasons are universal: an agent A's having a reason to *V* involves there being some fact *p* such that for any agent in circumstances relevantly similar to A's, that fact (or a relevantly similar fact) would be a reason for them to *V*. There are, in principle, no restrictions on what kinds of facts can constitute reasons in this way, but whenever some fact is a reason for some agent to perform some action, there is some general principle explaining why, which has this kind of universal form.

As we saw at the end of the last chapter, being rationally moved by a reason involves, for Scanlon, seeing it as a reason. Whether and how we should take the perceptual language of 'seeing' seriously and literally, seeing some consideration as a reason is a cognitive matter: it is an attitude that is correct or incorrect depending on whether the consideration in question is in fact a reason, independently of one's seeing it as such. No extra or prior motivational state needs to be added in order for a rational agent to be moved by something they see as a reason; rational agents are simply, as such, disposed to be moved by things they recognise as reasons. Scanlon, then, rejects the 'Humean' view that some prior motivational state, perhaps more specifically a desire, is always necessary for a belief or other cognitive state to motivate.[71] Moreover, he argues that 'what is generally called a desire involves having a tendency to see something as a reason' (Scanlon, 1998, p. 39), and that ordinary desires are not independent sources of motivation. The motivational force even of 'unmotivated' desires[72]—that is, desires that are not 'derived' instrumentally from further desires or aims—is, on Scanlon's view, to be understood in terms of the agent's seeing something as a reason. Someone who is thirsty is motivated to drink, for instance, because they take the facts that drinking would relieve the unpleasant sensations characteristic of thirst and that it would feel pleasant as reasons for drinking. Desire is not a state that motivates, but a state of being motivated by an apparent reason.[73]

Any putative desire that does not involve seeing something as a reason is, on Scanlon's view, deviant. He presses this point using Warren Quinn's famous 'radio man' example (Quinn, 1994). The example of a man who is disposed to turn on every radio he sees, but who sees nothing good or desirable in turning on radios, is not turning them on in order to listen to music or to distract or comfort himself or for any other intelligibly desirable aim, 'fails to capture something essential in the most common cases of desire', namely 'having a

[70]. What follows is a very brief summary of some of the ideas presented in (Scanlon, 1998, Chapter 1).

[71]. See (Smith, 1987) for a prominent example of such a view.

[72]. See (Nagel, 1978).

[73]. Compare (Alvarez, 2010; Dancy, 2000). Although the discussion in this chapter will focus on Scanlon's view, the arguments I make about desire apply just as much, in my view, to 'non-cognitivists' such as (M. Schroeder, 2007), who 'background' desire. Schroeder agrees with Scanlon that desiring involves seeing something as a worldly reason—he simply explains the latter in terms of the former rather than vice versa.

tendency to see something good or desirable about it' (Scanlon, 1998, p. 38). Unsurprisingly, given this picture of desire, Scanlon also holds that desires do not normally generate or provide reasons for action in themselves. The support for this is the same as that for thinking that an agent's beliefs do not generate or provide or constitute reasons: the reason, if there is one, is the fact which the agent's desire or belief represents as obtaining or as being a reason. If someone desires to do something that they have no worldly reason to do, their desire simply misrepresents what reasons they have.

On the Scanlonian picture, then, the idiosyncrasy introduced by desire is basically epistemic: insofar as an agent's desire must be invoked to explain their doing something that they lacked worldly reason to do, normally we understand their action in virtue of recognising how, in desiring as they did, some consideration seemed to them to be a reason even though it was not.

3.1.2 Hampshire's example

Scanlon nicely summarises his view about the connection between desire and reasons in a later chapter of *What We Owe to Each Other*:

> [T]he fact that I desire something does not itself provide me with a reason to pursue it. Being an object of a rational or 'informed' desire may be correlated with the presence of such reasons, but these reasons are provided not by this hypothetical desire, but by the considerations that would give rise to it, or make it 'rational.' (Scanlon, 1998, p. 98)

Stuart Hampshire, quoting this passage, in his review of Scanlon's book, suggests that this 'blurs a necessary distinction, which we all habitually recognize.' Hampshire illustrates his point with a memorable example:

> [A]n amateur collector of Italian Renaissance bronzes 'falls in love', as we say, with a particular bronze sculpture on sale, and feels that he must have it in his collection, even though, being of the wrong period and of doubtful provenance, it adds nothing to the distinction or value of his collection. The intensity of his desire is the reason that he would give for spending so much money and his justification also, and not, for instance, any further calculation or thought.
>
> ... My collector may certainly not subscribe to a universalizable principle that anyone is permitted to buy anything that he 'falls in love' with. He makes no universalizable claim. His entire feeling is directed toward this object here and now, and this feeling is his only 'justification.' (Hampshire, 1999)

Hampshire suggests that, although the example is quite unusual, the phenomenon it illustrates is familiar and commonplace:

> We surely all have such immediate experiences, such enchantments, even if they are more often concerned with food, drink, sexual desire, or with particular localities and memories than with bronzes. (Hampshire, 1999)

Because of the importance for the argument of Scanlon's book of the idea of 'having a reason that can be anyone's reason', Hampshire says, Scanlon sweeps such experiences under the carpet.

Hampshire does not develop this argument in any further detail. After all, it appears in a review of Scanlon's book as a whole, and the discussion of desire is just one part of that book, albeit a significant one. However, the argument seems to me to contain an important truth, and I think it is worth drawing out. While some of the panache of Hampshire's presentation will inevitably be lost in the process, I hope it will help to bring out the force of his point.

3.1.2.1 Clarifying the example

In Chapter 1, I characterised the special explanatory character of rationalisations by saying that a rationalisation explains the action from the agent's point of view, in that it shows the point that the agent saw in their own action and thus reveals something of the agent's self-conscious understanding of why they are doing what they are doing. In the first two chapters we investigated rationalisations which make sense of the agent's action by showing how, from the agent's point of view, their action was taken on the basis of some worldly reason: some fact that showed acting in the relevant way to be in some respect good or desirable or worthwhile. On a Scanlonian cognitivist picture, this is the only kind of rationalisation there is, in part because of the universality of reasons. The challenge that Hampshire's example presents to such a picture is just that while, one the one hand, the collector's action makes perfect sense to us, and we can see how from his perspective there was a perfectly intelligible point to his buying the bronze, there was apparently no adequate worldly, universal reason for him to buy it. If that is correct then the cognitivist view suggests that he must either have acted on the basis of a merely apparent reason—that is, he must have acted in ignorance or under a misapprehension about what worldly reasons he had—or he must have acted irrationally. Aside from defending a philosophical theory about the nature of reasons and rationalisation, though, it is not clear that we have any reason to characterise the collector's action or his perspective on his action in either of these ways. The collector, we can imagine, is a true connoisseur: he knows the value of the piece and knows that, in itself, it is not worth the money. If he did not feel about it the way he does, we would find his buying it unintelligible, but knowing that he has 'fallen in love' with it, his action is not unintelligible at all. Hampshire's claim, that the collector's desire itself is what gives him a reason to acquire the bronze, thus seems the right description of the case.

To make the example a bit sharper, let us suppose that the collector has to choose between two different bronzes. He has gone to a certain seller with a certain budget and has committed to buying something or other. Having looked at everything else the seller has to offer, he has identified one bronze that he judges to be worth the money. There is just one piece left for him to look at. When he sees it, he recognises its mediocrity straight away, as well as its inappropriateness for his own collection. Objectively speaking, it is not worth having, and the seller is asking far too much for it—as much, indeed, as the one decent

option that the collector already picked out. He would certainly not judge that anyone else in his situation would have a good reason to buy this piece. And yet, something about this piece speaks to him. He sees that it is trashy and kitsch, but he loves it. He feels he has to have it, and so he buys it, and the only reason he can give for doing so is the strength of his desire. He has no further reason to buy it, and yet his choice makes perfect sense to him. If we disregard his desire as a reason, we will not be able to understand this.

As Hampshire tells the story, the bronze 'adds nothing to the distinction or value of [the agent's] collection.' One way to read this is as saying that the bronze is utterly worthless, wholly without value, and such a reading might lead to an initial, and I think misguided, qualm with the example. The qualm is that the bronze must have *some* value, simply because it is wholly obscure what a *completely* valueless sculpture could be like. As I understand the force and point of Hampshire's example, though, it does not depend on the idea that the object of his desire is utterly without value, or that there could be no worldly reason for wanting this particular object. The cognitivist idea that an agent's going after some object of desire is only intelligible insofar as we can understand the agent as taking themselves to have some worldly reason to go after that object should not, I think, be taken to give us an absolute division between intelligible and unintelligible objects of pursuit, namely those which possess value and those which lack it. Again, the point is about the agent's perspective; it is about what they want the object *for*. It is also, in Hampshire's example, about the 'balance' of the agent's reasons. What is crucial for the example is not that the collector's only reason could be his desire because there is no other conceivable reason for buying this worthless object; what is crucial is that whatever worth the bronze has in itself, it is not, and could not be mistaken by the collector to be, enough to justify the cost of acquiring it. He nonetheless takes himself to be justified in acquiring the bronze. Hence it seems that there must be some other reason at play, and the obvious candidate is his desire.

A second point I think we should agree upon concerns the collector's relation, as we might say, to his own desire. Cognitivists about rationalisation often acknowledge a certain 'indirect' way, consistent with the cognitivist picture, in which desires can generate reasons for action.[74] Even a 'desire' that one experiences as a compulsion or mere urge, like the obsessive-compulsive's urge to wash her hands, or the committed dieter's craving for calorific food, can sometimes provide a respectable *worldly* reason to do what it is a desire to do. Because such a desire is typically experienced as unpleasant and distracting, and assuaging the desire will typically cause it to subside, the fact that washing one's hands will satisfy one's desire to wash one's hands can be a worldly reason to wash one's hands: it shows that washing one's hands will be worthwhile in that it will alleviate one's discomfort. There is nothing special here about the fact that the condition to be alleviated is a desire and that the way it is alleviated is through satisfaction. The point the agent sees in taking the relevant action, and the description under which it is rationalised, is just that of alleviating an unpleasant condition.

[74]. This point is made by, among others, (Alvarez, 2010; Parfit, 2011; Raz, 2000).

It is certainly correct that a desire can provide the desirer with a worldly reason in this way. The collector, though, does not buy the bronze in order to alleviate the discomfort of his longing for it. Neither need we think of him as, for example, taking his desire as evidence that the piece really does have some subtle but significant artistic value. He simply takes a shine to the piece, and indulges himself by buying it. This, just in itself, is perfectly intelligible. He is not fleeced by the seller into thinking that the piece is actually very good, he has not decided that kitsch is going to be the next big thing and that he should invest early. He just finds himself attracted to what is from the universal perspective the worse option, knowing that it is, objectively speaking, the worse option; he chooses it because of his attraction to it, and understands his own action in the light of that attraction.

3.1.2.2 Attraction and reasons

There is a weaker and a stronger way to interpret Hampshire's claim about the collector's desire. The stronger is that the desire need not be a 'response' to any apparent worldly reason at all. What it is about the bronze that attracts the collector may, on this reading, be entirely obscure even to him, and this need not undermine the desire's force as a reason for him to buy it. Call this the *non-cognitivist* reading.[75] The weaker—call it the *weak cognitivist* reading—still insists that the desire must, if it is to make the collector's action intelligible, involve his seeing something as a worldly reason, but that the strength of the desire, being disproportionate to the strength of the apparent reason, generates an 'extra' reason for buying the bronze.

Much of what I want to argue would be supported just as well by the weak cognitivist reading. My main argument will require only that the reasons or apparent reasons provided by how things stood from the collector's point of view cannot make sense of his doing what he does, and that to understand his action from his point of view we must recognise the role of his subjective response to the inferior bronze. That bronze's value was not in itself a good enough reason to buy it, and the collector did not take it to be. His reason, as Hampshire says, was his desire, or the strength of his desire. Even if we understand the collector's desire as involving some representation of a worldly reason, as long as the strength of the desire is not reducible to the apparent strength of that reason, we can understand the strength of his desire as itself providing a reason that is indispensable to our understanding of his action. So even if we see the collector's desire as involving his seeing something as a worldly reason, the example might still be used to show that the rationalising force of desires is not always reducible to that of apparent worldly reasons. Nonetheless, I think the non-cognitivist reading, on which the desire need not involve a representation of a worldly reason at all, more attractive.

The best case for thinking that the collector's desire must be a response to some reason is that unless he wants the bronze for a reason, his desire will be unintelligible, hence an 'alien

[75]. This should not be confused with (meta-)ethical non-cognitivism. As I have already explained, I take Hampshire's example to pose a problem for (for instance) forms of expressivism and quasi-realism that 'background' desire rather than allowing it to figure in an agent's self-understanding as a reason.

urge' (like the radio man's urge to turn on radios) hence unsuitable to make sense of his action. I will in later chapters argue that the first move of this argument is mistaken: the former's being based on an independent reason is not the only way to distinguish ordinary desires from alien urges. If this is right, then we are not compelled by such considerations to adopt the non-cognitivist account.

The basic advantage of the non-cognitivist reading is that it more realistically captures the kind of experience that Hampshire's example illustrates. The collector need not be able to say why the inferior bronze attracts him as it does, or what it is about it that appeals to him so much. He might try to do so, but it is perfectly conceivable that in doing so he is casting about, speculating, forming hypotheses. This is quite unlike the situation of the agent who self-consciously acts for a worldly reason, and who, in acting, understands their own action in terms of the reason for which they act. As Richard Wollheim observes, bringing out what it is about the object of our attraction that attracts us can be a real achievement (Wollheim, 1999, p. 15). In this, the kind of intrinsic attraction illustrated by Hampshire's example differs from intentional action. Knowing why you are doing what you are doing is not an achievement in the same way.

This point connects with something that I will suggest is an important feature of the collector's desire, namely its particularity. Experiencing this particular bronze, the collector finds himself attracted to *it*. If he were attracted to it in virtue of taking himself to have some worldly reason to acquire it, the reason in question would consist in the bronze's possessing some property or feature that could in principle be instantiated by another bronze. We would presumably, on a cognitivist picture of desire, expect the collector to be attracted in the same way to such an alternative. It seems to me, though, that we have a conception of a kind of attraction or desire that we do not expect to work in this way—a state in which a person simply fixates upon a particular object. This idea will be explored further in the discussion of love in Chapter 6.

Something that might seem to favour a weak cognitivist reading of Hampshire's example is the thought that there must be some apparent reason for him to buy the inferior bronze that does not equally apply to the superior one, some reason instantiated by the former but not the latter. Otherwise, the collector's differential attraction, insofar as it was based on an apparent reason, would surely be irrational. Suppose, for example, that you are offered two wads of cash, one amounting to £200, the other £210. It would be bizarre for you to take the smaller wad and to give as your reason for doing so that you 'fell in love' with it. This, we might think, suggests that, if the collector's desire is to be intelligible, we must imagine that it is based on *some* reason to favour it. Perhaps it instantiates some artistic value to a greater degree than the alternative, even though the alternative instantiates more values to a greater degree.

While I agree that the cash example would strike us as bizarre, it is not clear that the explanation of this must make appeal to apparent worldly reasons. Our puzzlement at the cash example might be explained by something else, such as our general understanding of

what kinds of things tend to intrinsically attract people.[76] It may be a part of our understanding of attraction that there has to be something *distinctive* to attract a person to an object, and we may have some ideas about what this distinctiveness has to be like. In particular, given that the attraction under consideration is a kind of experience, and that we are imagining it being formed on the occasion of an encounter with the object, we might think that the object of attraction needs to be distinctive in a way that might figure in the subject's experience, that might catch their attention. The cash example is a case in which the agent is presented with two options which differ only in respect of cash value, and it is just odd to think that a pile of cash might 'speak to' someone just because it is a little smaller than another one. It is hard to imagine what such a desire could be like. On the other hand, there are all sorts of things that we do understand people's being intrinsically attracted to. Our puzzlement at the cash case might just be that it is hard to see what the psychological story about the agent's attraction *could* be.

Something else that might be feeding into our judgements about the cash example is this. Accepting that being attracted to something can give one a reason to choose it does not mean that attraction is always a good enough reason for choosing it. In the cash example, it is hard to see how it could be much of a good reason. After all, the agent is presumably taking the cash in order to spend or save it, not just to have and to treasure. Even if we could make sense of the idea of one wad of cash rather than another catching one's eye, the main point of acquiring cash is for its monetary value, and intrinsic attraction is not the right kind of reason to feed into a decision properly governed by the kind of choice value that guides economic decisions such as this. Our judgements might be a little different if we imagined the case such that, for the agent making the choice, the monetary value of the respective wads is not a very significant consideration. Suppose for instance that these wads are being offered to a billionaire, and she just happens to take a shine to the £200 wad—after all, 200 is such a nice round number, and there is something about the way that the notes are rumpled just so .... Given the vanishingly small marginal value of the extra £10 to such a person, perhaps *her* whim *could* be enough to rationalise her choosing the smaller wad of cash.

I conclude that the attractions of the non-cognitivist reading of Hampshire's example are not obviously outweighed by any benefits in the weak cognitivist reading. I will henceforth pursue the non-cognitivist reading.

3.2 Desire as 'tipping the balance'

Before considering possible cognitivist responses to the challenge posed by Hampshire's example, I want, in this section, to compare that challenge with a similar argument for a similar conclusion, presented by Ruth Chang.[77] Like me, Chang argues for the view that desires can provide reasons for action, and does so on the basis of an example of forced choice, claiming that we need to understand the desire of the agent in the example as a reason in order to make sense of the rationality of that agent's action. I believe that the

[76]. See (Yao, forthcoming) for a view along these lines.
[77]. In (Chang, 2011).

argument I have presented, based on Hampshire's example, is in certain ways more compelling than Chang's argument, for reasons I will explain. The key difference between my argument and Chang's is that hers is based primarily on examples in which the worldly reasons in favour of each of the options open to the agent are 'evenly matched', and where each option is therefore rationally 'eligible'. Chang then claims that if we imagine the agent to form a desire for one of the options rather than the other, we can see that it would be *irrational* for the agent to then go after the other, non-desired option instead. The only way to make sense of this, she suggests, is to see the agent's desire as generating an extra reason, such that the agent has most reason to go for the desired option. Chang's argument differs from the present one, then, in that the desire in her case is supposed to make rationally compulsory an action that would otherwise be merely optional, whereas in Hampshire's case the desire makes sense of the agent's taking an action that would without the desire be rationally unintelligible.

3.2.1 Chang's argument

Chang's basic example is a familiar one:

> Consider Buridan's famous ass, poised between two equidistant and qualitatively identical bales of hay. There are, by hypothesis, no independent [worldly] reasons for him to eat the one bale rather than the other. Now suppose that he 'feels like' the hay on the left, not because it is to the left or for any other feature of it—he just wants *that* bale. (Chang, 2011, p. 80)

If in this situation the ass goes for the bale on the *right*—that is, the bale that is *not* the one to which he is attracted—Chang argues that he would 'surely … not be doing what he has most reason to do' (Chang, 2011, p. 80). What the ass has most reason to do, Chang claims, is to go for the bale on the left—the one he 'feels like' going for. Chang considers various ways in which a cognitivist might try to account for the ass's reason to go for the bale on the right, but finds none of them satisfactory, and concludes that therefore 'feeling like it' can rationalise an agent's going for one of two relevantly identical alternatives. She mentions other cases in which 'feeling like it' might play a similar role, such as choosing between cans of soup in the supermarket, choosing which of three identical slices of beef to eat first from one's plate, and so on. After that, she suggests that her conclusion can be generalised to apply also to cases in which the options are not relevantly identical but where one's reasons are still evenly matched. Then, finally, she suggests that, given that 'feeling like it' can make a difference to what one has most reason to do in these kinds of cases, there is no reason to think that it could not also tip the scales against the balance of worldly reasons, so that in some cases 'feeling like' the worse option can make it the case that that is the option one has most reason to choose.

This is to say that Chang's argument essentially *ends* where Hampshire's begins, on cases where a desire 'tips the scales'. Her thought seems to be that evenly-matched cases are the thin end of the wedge, and that once we have got desire-based reasons in there, there is no good reason to deny their existence in other cases or to deny that they might sometimes

make an 'ineligible' option 'eligible'. In a way, the difference between her argument and mine could be seen as a mere difference of strategy. Nonetheless, Chang's strategy seems to me to be the weaker, because it is much easier for the cognitivist to respond by simply rejecting her description of her example.

The first thing to note is that in the case of the ass, if he does choose the bale on the left (the one that he is attracted to), we do not *need* to appeal to his 'feeling like' going for that one in order to make his choice intelligible.[78] That action is perfectly adequately supported by objective, desire-independent, worldly reasons: the deliciousness and nutritiousness of the hay, or whatever it may be. Those reasons are good ones, recognised as such by the ass, and are not defeated by any other reasons. Of course, there is equally good reason to go for the bale on the right instead. The ass has conclusive reason, we might say, to go for one of the two bales, but he can satisfy this by going for either one or the other, and either choice would be perfectly intelligible from the ass's point of view. If the ass acts on his desire for the bale on the left, then, it is not necessary to have his desire explicitly in view in order to see his choice as making sense from his point of view. So the case is easily accommodated by the cognitivist picture, on which it is the agent's perspective on worldly reasons that basically makes sense of their actions. The desire itself adds nothing.

This (cognitivist) view of choice between eligible alternatives is nicely expressed by Joseph Raz:

> In these cases one understands (or thinks one does) what renders the action eligible. But one also understands ... that incompatible alternatives are also eligible, and not inferior to this action. It follows that one cannot understand from the inside one's preference for this particular action. ... Reason, so to speak, has exhausted itself. One cannot explain one's choice from the inside for there is no inside story to tell on that point. (Raz, 2000, p. 38)

In this respect Chang's example contrasts strongly with Hampshire's. In the latter case, the action would make *no* sense without the desire, because the worldly reasons so clearly favour doing something else and so clearly disfavour doing what the agent does. If the collector was not attracted to that bronze in the way that he is, his buying it would be deeply strange, even from his own point of view. This is what puts pressure on us to acknowledge that it is the collector's attraction to the bronze which makes his action intelligible.

Because of this, Chang's argument rests on the conviction that, were the ass to go for the bale on the right—the option he doesn't 'feel like'—he would be acting irrationally, against reason. Chang's argument is that because the worldly reasons are evenly matched, there must be some further reason that tips the balance in favour of the left bale so that if the ass went to the right he would be doing something he has most reason not to do. If he is doing other than what he has most reason to do, then it must be his desire that makes the difference. Chang's argument, then, rests crucially on a judgement to the effect that there would be something wrong with the ass's acting against his inclination, whereas Hampshire's rests on

[78]. I am imagining, for the sake of Chang's argument, that the ass has a basically human rational psychology. If you find this too silly, just think of him as a human with a predilection for hay.

the thought that there is, so to speak, something *right* about the collector's acting on his desire. The judgement that there would be something wrong with the ass's going to the right is hard to assess. There would, perhaps, be something odd about acting against inclination in this way when there is no reason to do so (no reason, that is, to act against inclination: there is perfectly good reason to go for that bale of hay), but is this oddness a matter of irrationality or unintelligibility? Or is it just the oddness of the unexpected and unexplained?

3.2.2 Picking and choosing

One way to frame this issue is with the distinction between 'picking' and 'choosing'. Ullmann-Margalit and Morgenbesser (henceforth U&M) explain this distinction as follows:

> We speak of *choosing* among alternatives when the act of taking (doing) one of them is determined by the differences in one's preferences over them. When preferences are completely symmetrical, where one is indifferent with regard to the alternatives, we shall refer to the act of taking (doing) one of them as an act of *picking*. (Ullmann-Margalit & Morgenbesser, 1977, p. 757)

We naturally understand 'Buridan's ass'-type cases as cases of picking. Chang wants to claim, in effect, that the ass's preferential attraction to one of the bales turns it into a case of choosing, and her argument for this is based on the claim that it would be irrational for the ass to act against that preference. There might, however, be ways to explain away this appearance and maintain that the case is one of mere picking.

One of the challenges that picking cases raise is: How do we pick? There are both ancient and modern takes on this question. Buridan's ass has traditionally been used in discussions of freedom of will, wherein the key feature of the case is that the ass's opting for one or the other bale is undetermined, the idea being that action in such a choice situation requires a special kind of freedom of will, the 'liberty of indifference' (Rescher, 2009). That acting in such cases requires a special free act of unconditioned will might seem unlikely if we understand the symmetry at issue in the cases as being simply a symmetry of the reasons for choosing one or the other option; it could, after all, be that something other than a reason, such as some subpersonal neurological event, determines which we go for. Nonetheless, picking cases do, from the agent's point of view, seem to present a difficulty. One needs to decide, but nothing tells one what to decide upon.

U&M suggest that what actually happens in such situations is that we simply have the ability to randomly pick.[79] Moreover, they offer a suggestion as to how the picking mechanism might work:

> [In a picking situation, you] haphazardly focus your *attention* on some one of the available alternatives. Once you do that, however, then—by hypothesis—none of the other alternatives attracts you more, and there is no room for qualms or second thoughts. (Ullmann-Margalit & Morgenbesser, 1977, p. 774)

[79]. They intend this claim in a way compatible with determinism. The idea is roughly that from our point of view as agents we can and do pick randomly.

Applying this suggestion to Chang's example, we might say this. The ass's inclination to go for the bale on the left is a matter of his having his attention arbitrarily focused on that one. That's the one he is attending to; it's a tasty bale; it's there. If he then goes to the right, his action is not made irrational by the balance of first-order reasons, since the first-order reasons for going to the right are just as strong as those for going to the left. However, there might seem to be something irrational about his arbitrarily resisting his inclination to go to the left, given that that inclination solves his picking problem. There is at least something odd, perhaps even something perverse, about what he does. We might even suggest that the risk of being paralysed by choice in symmetrical cases gives us a second-order reason[80] to follow any inclination towards one or the other option, so that the ass really does act against reason.

If we understand Chang's argument in this way, though, it can only establish quite a limited role for desire, because its conclusion is strictly limited to picking cases. The picture that results is one on which spontaneous attraction can settle for an agent which of a range of indifferent alternatives to go for. The desire provides a reason just because of its utility as a way of avoiding indecision, nothing more. The argument certainly does not show that that desire can tip the balance of reason to favour an option that would otherwise be disfavoured.

Chang might well insist that this reading misunderstands her argument. To refer to the ass's orientation towards the bale on the left as an 'inclination', something that might be understood along U&M's lines as the ass's merely attending more to that bale, is to misdescribe her case. Chang is quite explicit that what she is concerned with is a specific form of desire—what she calls *affective* desire. This is clearly meant to be something like the kind of felt desire or attraction that also figures in Hampshire's example. Chang's 'feeling like it' is meant to be a desire of this kind. So perhaps to see the force of Chang's argument, we just need to properly get inside the ass's head and see things from his point of view. We need to imagine not a mere inclination ('Well, I might as well go for that one') but a genuine *attraction* ('I just can't resist *that one*!'). On this reading, Chang is really redescribing the example of Buridan's ass for a very different purpose. The point is not about the possibility of picking and how attraction of some sort might play a role in this. The ass's desire for the bale on the left is meant not just to settle a picking problem but to transform what would be a mere picking situation into a genuine choosing situation. In desiring the bale on the left, the ass is no longer indifferent.

As U&M point out, there can be genuine choosing, because there can be real preference, between relevantly identical alternatives. They offer the example of children selecting from a plate full of identical sweets. While most adults would simply pick whichever one they happened upon, this kind of case, U&M observe, 'often poses [children] a serious and elaborate problem of choosing' (Ullmann-Margalit & Morgenbesser, 1977, p. 780). Moreover, once the child has made their determination, they would typically be unhappy for the sweet they chose to be replaced by one of the others. While this illustrates how taking the right attitude to a selection situation can transform a case of picking into one

[80]. For the idea of second-order reasons, see for example (Raz, 1986, 1999).

of choosing, the way it does so is not entirely friendly to Chang's argument. Consider the way U&M characterise the children's attitude:

> Children, we say, see differences where we do not see any, or take trifling differences to be relevant—that is, sufficient reasons (usually patently *ad hoc*) for preference. Indeed we generally regard it as a sign of growing up when a child stops 'behaving childishly' and is able to take a picking situation proper as just that .... (Ullmann-Margalit & Morgenbesser, 1977, p. 780)

In treating such situations as ones of choosing rather than mere picking, children behave irrationally. They think it matters which sweet they choose, when clearly it does not. They want to make sure that the sweet they select is the best one, even though the sweets are clearly all the same. The child forms a real preference for one of the options, but this preference is not a matter of 'feeling like' going for one option in recognition of the fact that no option is favoured by independent reasons. It is a preference based on the false belief that the option in question is the best.

It is not obviously implausible to say that we cannot really understand Chang's example without imputing some such illusion to the ass. For it is hard to see how someone could intelligibly fix in this way upon one of two options which they take to be in all relevant respects identical. There would, the cognitivist could insist, be something bizarre about such a preferential attraction.

There might seem to be real-life counterexamples to this claim. It sometimes happens, for example, that someone is 'preferentially' attracted to one of two identical twins. Such a desire is neither irrational nor unintelligible, and may well provide this person with a reason to seek further interaction with the twin to whom they are attracted. However, the case is not clearly a counterexample, since 'identical' twins are not identical in the relevant way. Even if two people look exactly the same, differences in their behaviour, their character, your particular interactions with them, and so on, may intelligibly generate a particular attraction to one of them and not the other. This need not be a matter of your arriving at any judgement to the effect that twin A is nicer or more attractive or in any other way 'better' than twin B and forming a desire on that basis, but perhaps it must be conceivable that there is *some* story to be told about what attracted you to twin A, or how you came to be attracted to twin A, that differentiates twin A from twin B. The conceivability of such a story seems to be missing in examples like Chang's, and as a result the idea of a strongly felt preference is hard to make sense of.

3.3 Possible responses

We have seen how the example of Hampshire's collector presents a challenge to the Scanlonian cognitivist. How might the cognitivist attempt to accommodate the example, and more generally the kind of phenomenon that it illustrates? There are a number of different strategies available, but none of them seems ultimately satisfactory. The first is just to deny that the collector buys the bronze for a reason. The second is to appeal to a

representational conception of desire such as Scanlon's 'desire in the directed-attention sense'. The third is to argue that there is after all some real or apparent worldly reason which rationalises the collector's action. The first two, it seems to me, describe the collector's action as irrational in a way which we need not see it as being. And as I will also argue, there is a good case for thinking that the kinds of reasons to which the third approach might most plausibly appeal are best understood as being based in the agent's desire, and so cannot themselves explain his motivation to buy the bronze.

3.3.1 Choosing for no reason

A proponent of the first kind of response to Hampshire's example will have to acknowledge that the collector's desire plays a role in explaining his action, but will insist that it does not *rationalise* his action. To say that his desire must be his reason for buying the bronze because it makes his buying it intelligible, this objector might say, is just to conflate explanatory and motivating reasons.

The objector is quite right to point out that from the fact that something explains an action we cannot immediately conclude that it is among the agent's reasons for acting. However, it is clear that the collector's desire does not 'make sense' of his action in just the kind of purely third-personal way that non-rationalising explanations might. If the explanation of his buying the bronze was that he was ignorant of its low value, or that he was drunk, or that there was something otherwise off with his brain function at the time of his buying it, none of these explanations would make sense of his action from his point of view. The collector's desire is not like this. The feelings that the bronze elicits in him are a crucial part of what makes sense of the action *for him*. It is his desire itself that makes spending so much money intelligible *from his point of view*; it is something to which he himself would appeal in giving an account of himself.

Admittedly, the way in which his feelings move him is not the same as the way in which considerations of an object's value, or of an action's justice, move an agent for whom they operate as reasons, and I agree that these are things that we should call 'reasons for action'. However, the question is not whether desires are reasons in exactly the same way as these sorts of things. The question is whether desires can play a role in making actions first-personally, rationally intelligible that cannot be reduced to the role of worldly reasons or of beliefs about worldly reasons. We might of course choose to stipulate that 'reason' is to refer only to considerations of objective desirability or choiceworthiness—to considerations that would be reasons for anyone in relevantly similar circumstances—but having made this stipulation we could still frame a question as to whether anything other than reasons (in this sense) or beliefs about reasons played a fundamental role in rationalising actions.

The idea that the collector does what he does for no reason suggests that there is no story to tell, from his point of view, about why he acts as he does. Recall the passage from Raz I quoted above. There, Raz moves straight from the thought that 'reason', by which he means (apparent) worldly reasons, does not favour the course of action that one chooses over the

alternatives, and that one's preference for that course one chooses is not explained by reasons concerning the value of taking the course one prefers, to the thought that 'there is no inside story to tell' as to why one takes that course of action. This seems to me a mistake. We can accept that one's apparent worldly reasons do not explain one's *desire*, and that there might thus be no available explanation of the desire itself from the agent's point of view, whilst rejecting the claim that there is therefore no account from the agent's perspective of their *action*. We can say this if we say that the desire, for which the agent may have no first-personally accessible reason, can itself be a reason for their action. Hampshire's collector offers a model of just this possibility.

3.3.2 'Desire in the directed-attention sense'

The second response to consider is one that claims that the collector, simply in desiring the bronze, in some sense takes there to be some worldly reason to acquire it, so that even if his desire does rationalise his action, our understanding of his action from his point of view is still fundamentally an understanding of action in terms of the agent's perspective on worldly reasons. One account of desire that suggests this sort of response is Scanlon's notion of 'desire in the directed-attention sense' (or 'directed-attention desire', as I will also call it). Scanlon defines this as follows:

> A person has a desire in the directed-attention sense that P if the thought of P keeps occurring to him or her in a favorable light, that is to say, if the person's attention is directed insistently toward considerations that present themselves as counting in favour of P.[81] (Scanlon, 1998, p. 39)

Scanlon introduces this idea to try to capture the kinds of ordinary desires that can conflict with our settled judgements about what reasons we have. Imagine, for example, that you are at home, alone, choosing a film to watch. Of the available options, you narrow it down to two: a light comedy or a very well-regarded 'serious' film, a real classic of 20th century cinema. You know you would enjoy watching either one, but also that you would be significantly more enriched by the latter, and you judge in light of this that the more serious film is the better choice: watching it would add more to your life than the alternative. Nonetheless, the idea of watching the light comedy just holds more immediate appeal: you know that *Persona* is the more worthwhile film, but it is just not not as tempting as *Dodgeball*. Although the desire to watch *Dodgeball* conflicts with your settled judgement about your reasons, it is nonetheless not something you experience as a 'mere urge' or an 'alien impulse'. It is a state in which watching *Dodgeball* seems especially attractive. Since Scanlon takes all intelligible motivation to be based in our taking things as reasons, he needs to explain the character of this attraction.

[81]. Scanlon assumes that desires all have propositional objects, which leads him here to use 'P' inconsistently. On the first occurrence, 'P' seems to be standing in for a sentence, whereas in the latter two it seems to stand in for a noun phrase or a gerund. We could correct this by inserting 'its being the case that' or 'making it the case that' before the second and third occurrences of 'P'. I will argue in the next chapter that the assumption that desires are all propositional attitudes ought to be rejected.

That your desire to watch the trashy film is stronger does not necessarily mean that that is what you actually do. You might resist temptation and do what you know is right. In talking about the strength of your desire, then, I do not mean how much it actually influences your actions, but something like how strongly it is felt. The idea of desire that we are interested in here is thus not that which is commonly meant by 'desire' in the philosophy of mind and action, which encompasses any and all motivational or 'conative' (as opposed to cognitive) states of mind. On that conception, the idea of doing anything other than what one most wants to do can come to seem quite mysterious.[82] The kind of case we are concerned with, though, involves a narrower notion of desire, one closer, in my view, to what the word 'desire' usually expresses when used outside of philosophy. A desire in this sense has real psychological presence for the desirer: desiring shapes the subject's experience; the desirer *feels* attracted by, or drawn to, the object of desire. The object occupies the desirer's consciousness, tempting them to pursue it.[83] Because not all *motivation* has the character of *desire* in this sense, there is no mystery, or at least much less mystery, about how one could do something other than what one desires most strongly to do. Since not all rational motivation has this character, though, it also raises a question for the cognitivist as to how we should understand such states of attraction.

This is what Scanlon's notion of desire in the directed-attention sense is meant to capture. Desires in the directed-attention sense are, as attentional phenomena, conscious. They shape the desirer's conscious experience. If you have a desire in the directed-attention sense to watch a comedy, the considerations that you take to count in favour of watching a comedy will occupy your consciousness—more so than the reasons for watching something more serious, assuming you lack a matching directed-attention desire to watch the serious film. Your wanting to watch the comedy is, on this view, a matter of your attention's being insistently drawn to such considerations that you take to speak in favour of watching a comedy, perhaps the most notable of which would be that it will be a pleasant and undemanding experience.

The notion of directed-attention desire, then, provides a nice cognitivist-friendly picture of the kind of attraction at issue. If you act on your desire to watch the comedy, what makes your action intelligible from your point of view, insofar as it is so, is whatever apparent reasons your attention was insistently drawn to—namely that it will be a pleasant and undemanding experience. As Scanlon puts it, 'the motivational force of these states lies in a tendency to see some consideration as a *reason*' (Scanlon, 1998, p. 40). Nevertheless, the idea of desire in the directed-attention sense seems to provide for a degree of slack between what one most wants to do and what one takes oneself to have most reason to do. For it seems entirely possible that you might be of the settled view that your reasons for watching the serious film are much more weighty, whilst nonetheless having your attention drawn more insistently to the considerations that seem to speak in favour of watching the low-brow comedy.

[82]. See for example (Davidson, 1980c).

[83]. Such a distinction between broader and narrower notions of desire is noted by, among others, (Davis, 1986; Schapiro, 2014; Schueler, 1995).

A crucial feature of the notion of desire in the directed-attention sense is that, while it can account for how a person's having a strong desire can explain their doing something other than what from their perspective they have most worldly reason to do, and without the desire's being experienced as nothing more than an alien external force, as in the radio man case, nonetheless these desires are not themselves reasons—they are, like beliefs, representations of reasons. When an agent has a directed-attention desire to do something that they have better reason not to do, this suggests that there is a kind of conflict in their state of mind. In particular, where the directed-attention desire presents considerations as reasons for *V*-ing where the agent believes that these considerations are not in fact reasons for *V*-ing, the agent will be at least somewhat irrational if they act on that directed-attention desire.

With the idea of directed-attention desire in view, might we use it to explain how the collector's desire makes sense of his buying the inferior bronze? A first question that arises is: what features of the bronze did the collector 'insistently' see as reasons for buying it? As I argued in Chapter 1, an adequate rationalising explanation, being an explanation that enables us to see what point the agent saw in doing what they did, will, insofar as it rationalises in terms of the agent's apparent reasons, tell us what those apparent reasons are. Merely being told that there exists *some* consideration that the agent took as a reason to do what they did does not rationalise their action—it does not enable us to understand the action from the agent's point of view. The directed-attention desire view suggests that, when we are told that the collector bought the bronze because he felt a strong attraction for it, this tells us that there were some considerations to which his attention was drawn as reasons for buying the bronze and that he was motivated by these apparent reasons to buy it. On this picture, the explanation in terms of the collector's desire does not tell us what it was that seemed to him to be a reason to buy the bronze. So the mere desire-ascription should not in itself give us a fully adequate rationalisation of his action.

A more serious problem, however, is that if we understand the explanation of the collector's action in terms of directed-attention desire, we seem to be forced to see his action as being in a certain respect irrational. A directed-attention desire that conflicts with the agent's settled assessment of the balance of reasons is, as Chang observes, somewhat analogous to a visual illusion. When we experience an illusion, such as that of a stick looking bent when half submerged in water, 'our attention is drawn to features of the stick that present themselves as reasons to judge that the stick is bent' (Chang, 2011, p. 65). Chang claims that this 'necessarily involves a tendency to judge that the stick is bent' (Chang, 2011, p. 65), as, presumably, a directed-attention desire is meant to necessarily involve a tendency to do what it is a desire to do. It is not clear to me that the second part of Chang's description of illusions is correct: although a stick half-submerged in water continues to look bent, it is not so obvious that, in someone who understands the nature of the illusion and knows the stick to be straight, this must really involve any inclination to judge that the stick *is* bent. However, this does not matter for present purposes, because the key respect in which directed-attention desire is analogous to visual illusion is this: if the agent acts or judges on

the basis of the apparent reasons to which their attention is insistently drawn, and so acts or judges against their own settled judgement, they will act or judge in a way that is, from their own point of view, irrational. They recognise themselves as acting or judging against the balance of reasons.

This seems on its face to be a misdescription of the collector's situation. At least, it seems so to me. Further pressure will be put on this account of the case in the next chapter, in which we will see that there is good reason to doubt whether we really understand desire in the kind of representational terms that Scanlon's account implies that we must. First, though, we need to consider the final strategy available to the cognitivist, in which it is accepted that the collector is acting for a good reason, but it is insisted that this reason is a worldly reason.

3.3.3 Desire as a worldly reason

If we cannot escape the thought that the collector buys the bronze for a perfectly good reason, we will have to say, if we are to maintain a Scanlonian picture, that he buys it for a good worldly reason. For Scanlon, this means that there will be a universalisable principle that explains why such a desire, or something that depends on the desire, is, for anyone in the collector's circumstances, a reason to to what it is a desire to do. As I have indicated, Hampshire thinks that this inserts something into the case that is not necessary for us to make sense of the collector's choice:

> My collector may certainly not subscribe to a universalizable principle that anyone is permitted to buy anything that he 'falls in love' with. He makes no universalizable claim. His entire feeling is directed toward this object here and now, and this feeling is his only 'justification'.(Hampshire, 1999)

Again, Hampshire's argument is not quite compelling as he presents it. We should not assume that the collector himself must be following some rule to the effect that he acquires just anything that takes his fancy in this way. This is not just because, as Hampshire claims, he need not endorse a universalisable principle which states that that is what anyone ought to do—perhaps he could follow such a rule for himself without thinking that anyone else should (although this might be objectionable on a picture like Scanlon's). It is also because of the more mundane fact that following such a rule would tend to make him rather a poor collector. His being a connoisseur, we would expect that his acquisitions would, in general, be guided by considerations of artistic merit, suitability for his collection, price, provenance and so on, and not just on what he happens to be whimsically drawn to. We can, and to make the best sense of the example should, view the collector's choice on this occasion as being somewhat out of the ordinary.

The cognitivist can and should respond to Hampshire by pointing out that the collector need not hold that *anyone* is permitted to buy *anything* that they 'fall in love' with. All we need to attribute to him is belief in a principle that implies that falling in love is sufficient reason for buying something like the bronze in the specific circumstances in which he buys it —circumstances in which, let us suppose, he has plenty of money, can spare the space, and so

on. However, if in general desires do not, as Scanlon insists, give people reason to do what they are desires to do, or to get what they are desires for, we will need some account of why *this* sort of desire constitutes such a reason on *this* sort of occasion. Conditions like the agent's having sufficient funds and space only suggest that certain 'defeaters' do not hold, where such defeaters might make the his action irrational despite the strength of his desire. What we want is an account of why the strength of his desire counts for anything at all, given that it surpasses what is warranted by any independent reasons in favour of getting the thing he wants. This will involve giving an account of how, in virtue of the collector's desiring the bronze, the bronze comes to fall under some suitably universal 'value'. The two candidates that suggest themselves are pleasure and well-being. I will consider these in turn.

The real challenge here will be for the cognitivist account of the collector's reasons to make sense of the collector's own perspective on his action. This, I will suggest, is the real stumbling block. While there are ways for the cognitivist to capture the idea that the collector's desire gives him a reason to buy the bronze, these accounts inevitably see the collector's rational motivation—his reason for buying the bronze, the point he sees in buying it—as consisting in a higher-order desire to satisfy his first-order desire. This seems to alienate the collector from his first-order desire in a way that does not reflect how we naturally imagine the case, or the way in which we understand our own actions when acting on similar kinds of non-reason-based motivations.

3.3.3.1 Reasons of pleasure

One value to which the cognitivist might appeal is pleasure. It is a familiar thought that there is some deep connection between desire and pleasure. Indeed, some well-known philosophical doctrines take this connection to be extremely intimate—the most notable among these being the view that all pleasure results from the satisfaction of desire,[84] and the 'psychological hedonist' view that all desire is aimed at pleasure. We need not endorse either of these extreme views, though, to recognise that there is some connection between pleasure and desire.

Pleasure is widely recognised as a kind of good, and it seems that the fact that doing something is or would be pleasant is in general a perfectly respectable worldly reason to do it.[85] One way for a cognitivist to try to explain the thought that the collector buys the bronze for a good reason, then, is to claim that he buys it for the pleasure it gives him.

There are a couple of different strategies that can be employed here. First, we might say that the collector takes pleasure in the bronze and thinks that by having the bronze in his collection he can prolong this pleasure. On this account, his desire for the bronze is itself simply a response to this (apparent) worldly reason for buying it: the (apparent) fact that it will give him pleasure. Second, we might say that the collector's desire for the bronze *generates* some kind of pleasure, so that while the desire is not itself a reason, he has a worldly,

[84]. See (Butler, 1729). For more contemporary (and somewhat different) reductions of pleasure in terms of desire, see (T. Schroeder, 2004); (Heathwood, 2006).

[85]. There are exceptions to this. In particular, some hold that pleasure taken in cruel or evil actions gives no reason to engage in such actions. We need not decide this issue for our purposes.

pleasure-based reason that depends causally on the desire. This latter strategy can be developed in two ways. One possibility is that the collector's desire for the bronze makes the bronze pleasant for him to look at, think about and so on, and he wants to prolong this pleasure. The other is that the collector desires to acquire the bronze and anticipates the pleasure of *satisfaction* that he will experience when he does so.[86] Again, the question is whether we can really make sense of the collector's understanding of his action in any of these ways. There seem to me to be specific problems with some of them as well as some general problems that apply to all of them. Before explaining those problems, though, I want to acknowledge what truth there might be in these suggestions.

One important fact is that desire does often terminate in the pleasure of satisfaction: generally speaking, it is pleasing to get what one wants. There are two ways to think of this. On a *semantic* conception of satisfaction, a desire specifies a way that the desirer wants the world to be and the desire is satisfied insofar as the world is that way—possibly unbeknown to the desirer. On this way of thinking, the pleasure of satisfaction could only be the pleasure of knowing or believing that one's desire is satisfied. On a *psychological* conception of satisfaction, the feeling of satisfaction is a kind of mental state which constitutes the last stage of a natural motivational cycle. This is nicely described by Mike Martin:

> First we have the onset of desire directed towards some outcome ... . This motivates an agent to bring about the outcome, and its presence in the agent's mind may lead to the accompanying emotions of anticipation or apprehension. When the action occurs and the agent is either successful or not, desire ceases. Where the action fails, the agent is left with regret. (Martin, 1999, p. 11)[87]

On this way of thinking, the desired outcome, when it happens, causes the desire to cease and leaves the agent with (pleasant) feelings of satisfaction.

It is plausible that these two ways of thinking about satisfaction correspond to two different kinds of desire, which may well be compresent in an agent's mind.[88] Significantly, in both cases the feelings of satisfaction, and hence the distinctive pleasures of satisfaction, depend on the desire itself. The object or action that was in fact the object of one's desire might be the kind of thing that could have given one pleasure even if one had not had a desire for it.

Consider an example. Someone who has a predilection for expensive Burgundy will probably be very pleased to receive a bottle as an unexpected gift. This is a pleasure that we might plausibly hold not to depend on a pre-existing desire. However, there is a different pleasure that could arise from the satisfaction of a strong desire for a particular wine. Suppose Karoline tries a 2012 Olivier Bernstein Clos de Vougeot at a tasting and falls in love with it. She forms a strong desire to have some in her cellar, but the wine is far too expensive

[86]. Some of these strategies are employed by (Parfit, 2011, pp. 67–8) in debunking subjectivist theories of reasons. Parfit is primarily concerned with 'normative' reasons. While Parfit argues that desires do not generate reasons, and on this I mean to disagree with him, there is reason to doubt that Parfit is a cognitivist in the present sense, since he seems happy to say that we sometimes intelligibly desire things for no reason.

[87]. See also (Wollheim, 1984, 1999).

[88]. Again see (Martin, 1999).

for her to justify buying any straight away. The desire persists and she harbours it for some time, during which time she might think about the wine often, or not very often at all. When it begins to seem that it might be feasible to acquire a bottle of the Bernstein, Karoline's excitement and anticipation grows, and when she finally gets it she experiences the distinctive delight of having a deep and lasting desire finally fulfilled. This seems a distinctive form of pleasure.

One way in which pleasure is connected with desire, then, is in satisfaction. But pleasure can also figure in a desire before the desire is satisfied. When one longs for something, one may spend a lot of time thinking about it and specifically *imagining* doing, having or getting what it is one wants to do, have or get. This imagining is typically pleasant, and one typically imagines the doing, having or whatever *as* pleasant. This can be an important part of how desire motivates. As Wollheim (Wollheim, 1984, p. 89) observes, the imagined pleasure of a desire's satisfaction can act as a 'lure', can 'erode' resistance on the agent's part to pursuing it, and can also reinforce the desire itself. Wollheim suggests that these characteristics of desire capture what truth there is in the doctrine of psychological hedonism.

The idea that desire is connected with pleasure in these ways might seem like grist to the cognitivist's mill. Thanks to these connections between desire and pleasure, we can acknowledge that desires sometimes generate reasons that are perfectly respectable worldly reasons, namely facts about what courses of action will bring pleasure to the desirer. As Parfit points out, while a reason of this kind might causally depend on our having such a desire, it

> would not *normatively* depend on our having this desire. If some act would give us pleasure, this fact gives us a reason to act in this way, whether or not this pleasure causally depends on our having some desire. (Parfit, 2011, p. 68)

This might be right, at least as far as reasons of pleasure go. However, there is something of a puzzle for the cognitivist who wishes to exploit this strategy to account for cases like Hampshire's collector. The Scanlonian cognitivist wants to say that intelligible, non-alienated motivation involves seeing something as a worldly reason. The problem here is that the worldly reason in question is only on the scene when the agent already has a desire, which is to say a motivation. Wollheim may be right that imagined pleasure tends to reinforce an agent's desire, but the pleasure of satisfaction, which depends on the desire itself, cannot be a reason upon which the desire is *based*. The motivation itself seems still to be deeply idiosyncratic in a way that cannot be captured just by thinking about the agent's perspective on their worldly reasons.

Now, if the question Hampshire's collector raised was simply whether the collector's action can be understood by the cognitivist as having a rationalisation, this might not be a very significant worry. Whether or not the original desire is rationally intelligible, we can understand it as generating worldly reasons that can figure in a rationalisation of his action. Remember, though, that the primary object of our investigation is a rational agent's self-understanding. This means that if the cognitivist account just sketched is to be adequate, the collector's own understanding of his action must be that he has an unintelligible desire that

he might, as it were, exploit for the pleasure of satisfying it. Not only does this seem to be a strangely alienated picture of the collector's relation to his own desire, it also seems likely that having this kind of attitude would undermine the very pleasure that is meant to be making his action intelligible. It is the desire for the bronze, not the desire for the feeling of satisfaction, out of which his feelings of satisfaction will arise, and it is from *this* desire's attaining its object that those feelings of satisfaction arise. Doing what he desires to do only because it will feel good, which seems to be what is required by the cognitivist view for him to do it rationally, seems the wrong way for him actually to attain that pleasure.

Perhaps these concerns can be addressed. Even if they cannot, there is still the first strategy I mentioned. According to this approach, the collector's pleasure does not depend on his desire. Rather, what happens in Hampshire's example is that the collector simply *very much enjoys* the inferior bronze, and the reason he wants to have it in his collection is so he can prolong this pleasure. Perhaps he anticipates all the time he will be able to spend looking at and enjoying the sculpture if he takes it home with him, and takes this to be what makes it worth the money.

Again, it does seem very plausible that, when we imagine the collector 'falling in love' with the bronze, we might imagine him as taking great pleasure in it. What is less clear is whether we need to understand the prospect of prolonging this pleasure as being what motivates him to buy the bronze. Once again, the question is whether this is a psychologically realistic description of the collector's understanding of his own action, and whether it can make sense of the rationality of what he does. One concern with this approach is that pleasure, as a value, seems to be relatively fungible. Of course, there are different kinds of pleasures, and we might not view these as simply interchangeable. The pleasure of cooking is different from the pleasure of eating, and each is different again from the pleasure of a job well done or the pleasure of riding one's bicycle. But for the collector, we need only consider trade-offs within a specific pleasure: the pleasure of viewing bronze sculptures. Since the collector presumably enjoys viewing good bronzes as much as he enjoys viewing this bad one, it is not clear why he should take the value of the pleasure to outweigh the considerations of value and money against which it is being weighed.

Another, deeper, problem for the suggestion that the collector buys the bronze for reasons of pleasure is that this fails to capture the character and role of his feeling for the bronze. Recall Hampshire's descriptions of the collector's state of mind: he 'falls in love' with the bronze; he feels he *must* have it in his collection; his entire feeling is directed towards this object here and now. We can perhaps imagine a case in which someone in the collector's position just enjoys this object so much that he decides to buy it for the pleasure it gives him, but it seems to me that if we do so we are imagining a significantly different case from the one that Hampshire describes. The recasting of his reasons in terms of pleasure, once again, fails to recognise the particularity of the collector's reason. This point will, I think, become clearer when we turn to look at love in Chapter 6. I will argue that love shares certain significant features with desire as I am suggesting we understand it in cases like the present

one. If there is a temptation to understand the collector's motivation in terms of the value of pleasure, I suspect that this temptation will be much weaker when we consider love.

3.3.3.2 Well-being

If pleasure cannot capture the collector's reasons, perhaps something else might. A plausible thought that could stand to justify the collector's choice is that, while it is not good always to do whatever one feels like, a life utterly devoid of such actions, actions in which one submits to one's fancies, is missing something. That is to say: living a good life, or at least one way of living a good life, involves occasionally indulging oneself in this kind of way. We might in this way try to subsume the collector's desire-generated reason under a broad but seemingly objective or worldly value, the value of well-being. Once again, there seems to be some truth here. Insofar as it is plausible to see the collector's desire as giving him a reason to buy the bronze, it seems plausible to think of his desire as making the bronze part of his well-being, part of the good life for him. Once again, though, the question is whether the cognitivist can really make adequate sense of this truth. The obvious worry is that while the claim about the constitution of a good life is true, its truth is explained by something about the character of these self-indulgent actions, whereas the kind of account a cognitivist would have to give will, we might suspect, necessarily distort the perfectly ordinary phenomenon of indulging a whim.

It might be helpful here to compare policies that could be adopted by someone responsible for governing another person's behaviour. Most parents, and perhaps all good parents, will occasionally, but not always, indulge their children's whims. That a parent should occasionally, but not always, indulge their child's whims seems like a reasonably good principle. At least one way of unpacking this idea does so without treating the child's whims as in themselves constituting reasons for the child to do or get whatever it is that they want. Parents, indulging their children's whims, might see themselves as allowing their children to act irrationally, and perhaps even, in a way, unintelligibly. That is, they might see the child as lacking any good reason to do what they want to do, but think that, for some other reason, it is good to allow one's children occasionally to behave in such a way. They might take it to be better for the child's psychological development not to have their whims always frustrated. That might breed resentment; or perhaps an occasional indulgence of one's irrational whims is an important part of learning self-control.

Taking such a paternalistic attitude to *oneself* is rather less familiar, but there are certainly cases that we are accustomed to imagining. Consider the obsessive–compulsive who very frequently experiences compulsive urges to wash her hands. She, like the parent, might see these urges as in themselves irrational (and in this case, because she herself experiences them, 'alien') but nonetheless think that it is better to occasionally give in to them. Perhaps she knows that if she tries always to resist her urges, she will inevitably backslide in a much more damaging way. When she washes her hands, she treats her 'desire'

as a reason to do so in virtue of a higher-order principle explaining why it is a reason to do so. It should be clear that the obsessive–compulsive's desire in this example explains her action, makes it intelligible, in a very different way from that in which Hampshire's collector's desire makes sense of his. On the most natural reading of that case, the collector is not 'alienated' from his desire; rather, he endorses or 'identifies' with it.

To make a case against the Scanlonian view, we need an understanding of Hampshire's example on which, although the collector's desire explains his buying the bronze in such a way that it makes sense to call the desire a reason, his action is not based on some higher-order principle that explains why the desire is a reason to take the relevant course of action. This might be taken to suggest that we have to think of the collector as being completely unreflective, and that in itself might seem implausible. Surely, as a rational, self-conscious agent, the collector will have some kind of reflective view about his desire. He is not, in buying the bronze, behaving like an animal. If we are to say that he is not alienated from that desire, though, it seems we will have to say that his higher-order attitude is, in some sense, an attitude of endorsement. This might seem to be all the material that the Scanlonian cognitivist needs to accommodate the case.

This would be a mistake. What a Scanlonian account of rational motivation requires is that an agent who acts intelligibly for some reason responds to an (apparent) fact that constitutes their (apparent) reason because of a principle that explains why the fact is a reason to perform that action. The collector might, for example, be proud of his (occasionally bad) taste and think it good to buy in ways that express that taste. Such an attitude, though, does not explain the way in which his taking a shine to the bronze gives him reason to buy it. He might buy the bronze in order to express his taste, and here at least part of his reason for buying it is that buying it will express his taste, expressing one's taste being in general a good thing. If you value individuality, for instance, you may think it desirable to express your taste even where your taste is less than perfect. This value presupposes, though, that you really do like, and want to have, the things you like: otherwise, the acquisition of those things would not be an authentic expression of your taste. Compare, for example, someone with purely exquisite taste who, learning that the likes of Pablo Picasso and Helmut Newton thought good taste the enemy of creativity, becomes concerned that his preferences make him dull and bourgeois and so pretends to some degree of trashiness. Such a person would be fooling himself if he saw himself as expressing his taste in buying tasteless objects: it would not really be his taste that he was expressing. By the same token, the possibility of buying something you like in order to express your taste seems to be parasitic on the possibility of simply buying something because you like it.

Buying something to show people (perhaps including yourself) what kinds of things you like and so what kind of a person you are is not the same as buying something just because you like it. Of course, you might buy something both because ('just' because) you like it *and* because in doing so you will express your unique, sophisticated, well-informed-yet-subversive taste. In doing that, though, you are just self-consciously *expressing* your desire (your taste);

you are buying the thing just because you like it, in the awareness that that is what you are doing and with the view that acting in such a way is a good kind of thing to do.

3.4 Loose ends

In this chapter, I have tried to motivate the thought that cognitivist accounts of rationalisation will struggle to make good sense of the way that desires sometimes figure in our understanding of our own actions. If these desires cannot be accommodated by the cognitivist approach, cases like that of Hampshire's collector seem to suggest that our reasons for acting are sometimes subjective or idiosyncratic in a way that beliefs are not, and that desire can introduce a kind of idiosyncrasy into rational action that is much deeper than the idiosyncrasy of a fallible perspective on what worldly reasons one has. They suggest that as well as having reasons anyone can share, we also have reasons that are essentially our own.

I considered a few different ways that the cognitivist might try to accommodate the phenomenon illustrated by Hampshire's example. While I raised some difficulties for each envisaged response, there is inevitably more to say in each case. In the remaining three chapters, then, I will seek to reinforce the case made in this chapter by investigating more deeply some of the issues that have arisen. First, in Chapter 4, we will consider whether desire might plausibly be understood as a representational state. Scanlon appeals to the idea of desire in the directed-attention sense to try to accommodate its idiosyncrasy, and this suggests that our understanding of the idiosyncrasy of desire is, like our understanding of the idiosyncrasy of belief, an understanding of a mental state as representational. I will argue that there is reason to doubt whether our fundamental understanding of the idiosyncrasy of desire construes it representationally. Another question that has arisen is that of the collector's relation to his own desire—or, more generally, the way an agent relates to a desire which they take to give them a reason to act but which they do not see as based on a worldly reason. This connects with concerns about alienation and identification which I will discuss in Chapter 5. Finally, in Chapter 6, I will examine another potential source of non-universal reasons, somewhat different from, and arguably more important than, the kind of whimsical desire illustrated by Hampshire's example. Love, while it raises some puzzles of its own, will also help to shed further light on some of the issues already raised.

Chapter 4

Desire as Representation and the Representation of Desire

4.1 Metarepresentation

Just as our account of perspectival rationalisation had to explain the idiosyncrasy of belief, an account of how desire rationalises action needs to explain the idiosyncrasy of desire. As we saw in the previous chapter, the Scanlonian cognitivist attempts to account for desire's idiosyncrasy in much the same way as the account given in Chapter 1 explained the idiosyncrasy of belief. Desire, on this account, involves representing some consideration as a reason to take some course of action. Desire, like belief, is idiosyncratic just in that it can *misrepresent* what worldly reasons the agent actually has. In such cases, understanding an action from the agent's point of view requires us to appreciate how things seemed from the agent's perspective and how, as it seemed to them, they had some reason to do what they did. On this picture, understanding actions either on the basis of idiosyncratic belief or on the basis of idiosyncratic desire is what is sometimes called a *metarepresentational* task: it involves thinking about (hence representing) a representational mental state as such. To believe is to represent some state of affairs as obtaining; to desire is to represent some consideration as a reason. Understanding someone's actions on the basis of a false belief involves appreciating that they took things to be a way that they were not, and thinking about what it would have made sense to do had they been correct (whilst at the same time remembering that they were not correct); understanding someone's actions on the basis of an idiosyncratic desire involves appreciating that they took some consideration to be a reason in a way that it was not, and thinking about how it would have made sense for them to act as they did had the consideration been a reason for the relevant action (whilst at the same time recognising that this consideration was not such a reason). In this chapter, we will see that there is reason to doubt whether understanding idiosyncratic desire really does require metarepresentation, and thus reason to doubt whether our fundamental understanding of the idiosyncrasy of desire sees desire as representational.

4.1.1 In what sense representational?

Before I go on, I want to briefly clarify what it is that I am investigating, namely the claim that desire is representational. It is fairly widely accepted in philosophy and in cognitive science that the mind is essentially representational, but there are different ways to interpret this claim. I am concerned with the thesis, as it applies to desire, only on a specific and quite strong interpretation. We can broadly distinguish two kinds of representational theory of a given mental state or of the mind as a whole, which I'll call analytical and empirical.[89]

An empirical representational theory holds that the best scientific theory of the mind, or of the state in question, will explain it at least partly in terms of representation. An empirical representational theory of desire might claim that desire consists in, or is grounded in, or is realised by neural 'representations' of certain kinds, which play certain roles in the functioning of the organism, in something like the way that cognitive neuroscientists explain an animal's spatial memory and navigational abilities in terms of so-called place cells 'representing' locations in the animal's environment. An analytical representational theory, on the other hand, says that it is part of the concept of the state in question that the state is representational. Hence on an analytic representational theory of desire, a full, mature understanding of the nature of desire must involve an appreciation of its representational nature.

Since my concern here is with the place of desire in our ordinary understanding of action, in the 'manifest image', so to speak, I am concerned only with the claims of an analytical representational theory. Moreover, I am here concerned with a specific kind of analytical thesis about desire, namely that desire is a propositional attitude: that a desire is a representation of a propositional content or a state of affairs or a set thereof. There could be other ways of thinking of desire as representational that do not claim this. We might, for instance, think that desires have nonconceptual, and hence nonpropositional, representational content. Or we might think that desires represent their objects but that the objects of desire are not, or at not all, propositions, such that when I desire a doughnut, for instance, I mentally represent *a doughnut*, or perhaps *eating a doughnut*, but I do not represent *that I will eat a doughnut in the near future*, or (as on a Scanlonian picture) *that the doughnut's tastiness is a reason to eat it*—or at least, my representing some such proposition is not fundamental to my desiring a doughnut.

As I will explain later, if desire is representational just in that desiring a doughnut involves mentally representing *a doughnut*, or *eating a doughnut*, then desire's representational nature will not be so central to understanding its idiosyncrasy as it is on the Scanlonian cognitivist account. So the real question will be whether desire is representational in the sense of being a propositional attitude. Much of the literature I will be discussing simply assumes that desire must really be representational in this sense. The major source of disagreement is over *when* children come to understand it as such. I will argue, however, that this literature does not give us any compelling reason to think that desire—at

[89]. Compare the distinction between 'conceptual functionalism' and 'psychofunctionalism' in (Block, 2007).

least, the kind of desire at issue in the case of Hampshire's collector—must be representational in that sense.

4.1.2 The development of metarepresentation

Interestingly, it seems that children only develop the capacity to metarepresent at a certain age. The canonical way to test whether children can understand representational mental states as such[90] is to see whether they can understand actions that flow from false belief, false belief being perhaps the clearest and most uncontroversial case of mental misrepresentation.[91] The classic example of this is what has come to be known as the *direct* (or *verbal*) *false belief test,*[92] which children only start to pass at around their fourth birthday.

In the classic version of the direct false belief test, subjects are presented with some version of the following scenario. Sally and Anne are in a room with two boxes. Sally puts her marble in one box and leaves the room. While Sally is out of the room, Anne moves Sally's marble to the other box. Sally re-enters the room. The critical test question is then put to the children: Where will Sally look for her marble? Wimmer and Perner (1983) found that children only began to give the correct answer to this question—that Sally will look in the first box, where she left the marble—after their fourth birthday, whereas children younger than four consistently indicated that Sally would look in the second box, where the marble actually is. Wimmer and Perner concluded that children only develop the ability to understand false belief around the age of four. Subsequent studies using variations of this paradigm have shown this finding to be remarkably stable and robust.[93]

The claim that a general metarepresentational capacity develops around the fourth birthday is not universally accepted. Some authors argue that there are other competences that also require metarepresentation, but which children develop before they pass the direct false belief test. For example, Wellman and Estes (1986) argue that three-year-olds' understanding of imagination shows that they can metarepresent, and Leslie (1987) makes a similar case regarding their appreciation of pretence. Understanding the difference between imagined situations and real ones, and showing some appreciation that the former are 'just in one's mind', is, these authors argue, a metarepresentational task: it requires one to appreciate that these situations are merely represented rather than real. However, it is not clear that this is right: the imaginary–real distinction need not necessarily be understood in representational terms. It could instead be that, as Perner (1991) argues, children at this age are working with a 'situation theory'. Under-fours, on this view, distinguish imaginary or

[90]. The phrase 'understand representational mental states as such' should not be taken to suggest that metarepresentation, in the sense at issue here, necessarily requires possession or application of the concept *representation*. What is intended is something more along the lines of an appreciation of the kinds of features that representational states distinctively possess, most notably, for present purposes, the potential to be false or incorrect.

[91]. The reason for focusing on *mis*representation should be obvious: children who show understanding of an action based on true belief might just be understanding the action in terms of how things really are.

[92]. The classic examples of this paradigm are (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983).

[93]. See (Wellman, Cross, & Watson, 2001) for a meta-analysis of over a decade's worth of research on this topic.

pretend *situations* from real ones, but this division of situations into real and imaginary does not require one to think of anyone as mentally representing the latter any more than that would be required by the division of situations into actual and possible. A person's imagining a situation can, within a situation theory, be understood relationally rather than representationally: to imagine a situation is simply to be appropriately related to an imaginary situation. What under-fours cannot do with this purely situational model is to appreciate the possibility of *mistaking* an unreal situation for a real one.[94] Children of the relevant age might thus have a conception of the mind as 'intentional', in a certain sense, but not as representational.

There is other evidence of early metarepresentation that cannot be addressed in the same way because it is taken specifically to demonstrate early understanding of false belief. Onishi and Baillargeon (2005), for example, found that children as young as 15 months showed surprise (based on measurement of looking times) when, in a scenario analogous to the Sally–Anne story, the 'Sally' character looked in the displaced object's new location rather than where she left it. The authors argue on this basis that even infants predict behaviour on the basis of agents' false beliefs. Other studies, including ones using other 'indirect' methods such as anticipatory pointing and anticipatory looking, provide further support for this view.[95] These findings are taken to show that even infants are capable of metarepresentation and that under-fours' difficulty with direct false belief tests is better explained by some experimental artefact.

These results seem to be fairly robust. However, they are puzzling. Not only explicit understanding of false belief, but a range of other apparently metarepresentational capacities, including understanding of pictorial representation, verbal representation and alternative naming, all seem to develop around the same age that children start to pass the direct false belief test.[96] It would be quite surprising, given this convergence, if children could in fact metarepresent at a much younger age. Fortunately, there are ways of explaining the findings of indirect tests that do not require us to ascribe genuine metarepresentational capacities to under-fours. One of the more compelling of these is developed by Ian Apperly and Stephen Butterfill.[97] Apperly and Butterfill (A&B) suggest a 'two systems' model of so-called 'mindreading' or 'theory of mind' abilities. Adult humans, on this account, have two distinct ways of thinking about mental states: as well as the powerful and flexible but slow and inefficient system of metarepresentational concepts, we have a relatively encapsulated, automatic, less flexible but much more efficient mental state-tracking 'module', which A&B call 'minimal theory of mind' (minimal ToM). A&B argue that what competence under-fours show in 'predicting' false-belief-based action can be explained by attributing to them such a minimal ToM module. There is therefore no reason to think that they can genuinely metarepresent, because minimal ToM is not best understood as representing

[94]. Compare the emphasis, for instance in (Dretske, 1994), on the idea that the possibility of misrepresenting is essential to genuine representation.

[95]. Much of this evidence is surveyed in (Baillargeon, Scott, & Bian, 2016).

[96]. See (Perner, 1991; Perner, Brandl, & Garnham, 2003; Perner, Rendl, & Garnham, 2007; Perner, Zauner, & Sprung, 2005).

[97]. (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013).

representational states as such. Rather, it tracks them by representing simplified, merely relational states that are, within a wide range of ordinary situations, coextensive with genuinely representational states like belief. The idea is that such a system could be employed by agents with limited cognitive resources (such as non-human animals, infants, and human adults under cognitive load) to track other agents' mental states within certain 'signature limits'—limits predicted by the fact that minimal ToM does not truly metarepresent. To give a sense of A&B's approach, I will briefly outline their treatment of young children's ability to track others' beliefs.

Metarepresentation is complicated in a number of ways. Understanding what someone believes might involve not only entertaining some possible state of affairs that does not or might not obtain, but also thinking about non-existent or never-existent objects, understanding and using quantified predications, informative statements of identity, opaque contexts and so on. This goes some way to explaining why metarepresentation is demanding. But how could children keep track of beliefs in a useful way without representing them as representations? A&B suggest that within limits this could be achieved indirectly, by tracking a kind of relational, non-representational, state that partially overlaps with belief. A&B call this state 'registration'.

Registration as A&B define it is a relation between an agent *a*, an object *o* and a location *l*. Roughly, *a* registers *o* at *l* just in case she most recently 'encountered' *o* (that is, had *o* in her perceptual field) at *l*. Registration is connected with action in two ways: first, success in goal-directed action where one's goal specifies a particular object depends on having correctly registered that object; second, when one acts in pursuit of a goal that specifies a particular object, one will act as if the object were in the location where one registered it.

It should be clear enough how registration overlaps with belief to a certain extent. In the Sally–Anne story, for instance, Sally registers her marble in the first box, where she left it. This predicts that, since the marble has moved, she will not successfully find it, and that she will look for it in the first box. However, ascriptions of registration are considerably simpler than ascriptions of belief, for instance in that ascriptions of registration do not introduce intensional contexts, and only actually-existing objects can be registered. A&B's explanation of the findings about children's precocious belief-tracking abilities is that children from an early age possess a relatively encapsulated and automatic system, minimal ToM, that tracks registrations. Passing the direct false belief test, however, requires later-developing capabilities like language and, perhaps, a general capacity for metarepresentation. It remains plausible, then, that genuine metarepresentational understanding only develops around the fourth birthday.

4.2 Children's early competence with desire

If the idiosyncrasy of desire were fundamentally a matter of (mis)representation, as cognitivists claim, then we would expect children to develop an understanding of idiosyncratic desire only around the same time that they develop the general capacity for

metarepresentation, and in particular only around the age that they start to pass the false belief test. Helpfully, some relevant empirical work has been done on the development of children's understanding of desire, and interestingly, children seem to demonstrate a fairly good grasp of the idiosyncrasy of desire long before they are capable of passing the false belief test. Children's understanding of desire and their understanding of belief do not develop in parallel. This, on its face, suggests that our understanding of the idiosyncrasy of desire is not fundamentally metarepresentational. If that is right, it would seem to cast further doubt on the cognitivist account of how desire-ascriptions rationalise action.

However, we should not be too hasty. Many authors on the topic of children's theory of mind capacities simply assume that a mature understanding of desire must be metarepresentational. The observed 'asymmetry' in the development of children's understanding of belief and desire therefore generates a debate about whether under-fours' appreciation of idiosyncratic desires shows that they are understanding these desires as representational or whether it can be explained in some other way, along the lines of Perner's explanation of younger children's grasp of imagination and pretence. Authors who endorse this latter kind of approach have also worked to devise tests intended to do for desire what the direct false belief test does for belief: to test children with a scenario which they could only properly understand metarepresentationally.

So these authors on children's theory of mind abilities agree with the Scanlonian cognitivist that desire is fundamentally understood in representational terms. It should be noted, though, that they each conceive of this representational character in somewhat different ways. The Scanlonian sees the idiosyncrasy of desire in essentially cognitive terms: desire is, or involves, a representation of some consideration as a reason. This is a representation that, like belief, can be true or false, and understanding action on the basis of idiosyncratic desire involves appreciating how for the agent something seemed to be a reason even though it was not. For the authors interested primarily in children's development of theory of mind, on the other hand, the assumption that desire is representational tends to be fostered by the deeper theoretical assumption that mental states, in particular belief and desire, are propositional attitudes, representational states with propositional content, differing only, or at least primarily, in their 'direction of fit'. On this picture, what a desire does is to represent a state of affairs not as actually obtaining, but as 'to be brought about'. Nonetheless, desire is supposed to be representational: the subject has, so to speak, a picture of the world in their head, and this is a picture that can fail to correspond accurately to the real world.[98] Despite this difference in detail, both accounts predict that mature understanding of desire is metarepresentational, and so both face a challenge from children's apparent early understanding of idiosyncratic desire. It is worth looking, then, at how researchers have attempted to demonstrate a metarepresentational understanding of desire in children.

Before we consider the different attempts to devise a desire analogue for the direct false belief test, we should consider some of the different explanations that have been offered for

[98]. See for instance (Gopnik & Wellman, 1992).

under-fours' apparent competence in predicting and explaining action on the basis of idiosyncratic desires. And before we do that, we need to have a look at what exactly it is that these children are capable of—what evidence they do show of an understanding of desire's idiosyncrasy.

In Repacholi and Gopnik (1997), an experimenter interacted with 14- and 18-month old children with a plate of broccoli and a plate of biscuits. In the test condition, the experimenter expressed pleasure on tasting some broccoli and expressed disgust upon tasting a biscuit. The experimenter then requested some food without specifying which ('Can you give me some more?'). Since the children themselves overwhelmingly preferred biscuits, we would expect children with no appreciation of idiosyncratic desire to offer the experimenter biscuits. While this is exactly what most 14-month-olds did, most of the 18-month-old group offered her broccoli, thus apparently showing sensitivity to her idiosyncratic preference for the food that they themselves found digusting.

In another study, Yuill (1984) found that three-year-old children displayed an understanding of the dependence of another's emotional state on the satisfaction or frustration of their desires. Here, children were presented with a story in which a boy wants to throw his ball to a girl, but the ball is caught instead by another boy. Three-year-olds correctly judged that the first boy would be sad as a result. (They also judged that he would be happy in the scenario where the girl catches the ball.)

Understanding of emotional reactions, as well as ability to predict action based on desire, was also tested in two-year-olds by Wellman and Woolley (1990). They presented children with three different scenarios:

- *Finds-Wanted*: The protagonist wants something that could be in one of two locations, searches in location 1 and finds what they wanted.
- *Finds-Nothing*: Identical to Finds-Wanted except that the desired object is not found in location 1.
- *Finds-Substitute*: Identical to Finds-Wanted except that the object found in location 1, while desirable, is not the object that the protagonist desired.

Children were asked questions both about how the protagonist would act and about his or her emotional reaction. The vast majority of children studied correctly predicted that the protagonist would continue searching for the desired object in the Finds-Nothing and Finds-Substitute scenarios, and that he or she would be happy in the Finds-Wanted scenario but unhappy in the Finds-Nothing and Finds-Substitute scenarios.

Each of these studies shows evidence of children under four, who should be incapable of metarepresentation if the developing capacity for metarepresentation is indeed what enables children over four to pass the direct false belief test, have some ability to comprehend, or at least to track and respond appropriately to, the desires and preferences of other people, even where those desires and preferences differ from their own. Either these younger children can metarepresent after all, or metarepresentation must not be required for the level of understanding that the children in the studies under discussion seem to display. An obvious question, then, is whether we can make sense of how these children are thinking without

crediting them with a capacity for metarepresentation—as, for instance, Perner's 'situation theory' did for children's understanding of pretence and imagination. Another question, in principle distinct from this, is whether the children who answered successfully in these studies might be doing so because they really do understand the idiosyncratic character of desire.

4.2.1 Interpretations of these findings

There are two main ways to accommodate these results without attributing a metarepresentational capacity to under-fours. On the first, we see children under four as simply applying a teleological schema of action explanation: they see people as acting in pursuit of objectively desirable goals. The idea here is in effect to see children as having at their disposal something like the worldly, but not the perspectival, form of rationalisation. The appearance of idiosyncrasy is accommodated in terms of idiosyncratic worldly reasons. The second strategy attributes to under-fours a genuine concept of desire, but one that is merely intentional or relational, rather than representational. On the assumption that the idiosyncrasy of desire really is the idiosyncrasy of representation, this approach will not seem to provide for a true understanding of the real source of idiosyncrasy in desire. However, if we do not assume that desire *must* be representational in the relevant sense, we should keep our minds open to the possibility that children under four really do understand something basic about the nature of desire even though they do not yet understand the nature of belief.

Repacholi and Gopnik's findings, it seems, can be fairly straightforwardly accommodated by the first, 'objectivist' approach. Their study shows that young children do not simply assume that everyone likes what they themselves like. 18-month-old infants do not, liking biscuits and disliking broccoli, simply think 'Biscuits are good; broccoli is bad'—they are not in this sense completely egocentric. By the time these infants reach 18 months, they have clearly acquired a more complex understanding of *something*; however, it may be that what they have acquired a more complex understanding of is just worldly reasons, or perhaps 'the good'. They might, perhaps, simply have come to appreciate that biscuits are good for some people and bad for others. Or, modelling their understanding in terms of states of affairs, it might be that they come to appreciate that while the state of affairs in which they themselves have biscuits is good, and the state of affairs in which they themselves have broccoli is bad, the state of affairs in which this other person (the experimenter) has broccoli is good and the state of affairs in which they have biscuits is bad. Because these are all distinct situations, there is no requirement that the infant be able to integrate conflicting or contradictory attitudes to one and the same state of affairs—something which would arguably constitute a metarepresentational task.

Of course, if the child does evaluate differently the state of affairs in which they themselves have broccoli and the state of affairs in which the experimenter has broccoli, there must presumably be some basis for this difference—they must see some salient difference between themselves and the experimenter. One possible basis for the evaluative

difference would of course be that the experimenter *desires* broccoli and not biscuits. If so, the children's judgement that broccoli is good for the experimenter would be based on an ascription of an idiosyncratic desire for broccoli together with the thought that what someone desires is good for them. It might be argued, however, that there is no obvious reason to think that the infants' differential evaluations must be based on a consideration of desire. They have, after all, seen the experimenter emote positively when tasting broccoli and negatively when tasting biscuits, and this might in itself be enough to mark the former situation as positive and the latter as negative. So long as the infant can appreciate the one expression as having positive valence and the other as having negative valence, there seems to be no need for them to infer anything about the experimenter's subjective state of mind or her 'point of view' on broccoli and biscuits. On the other hand, we have not yet seen any compelling reason not to attribute a genuine understanding of desire to these children.

The children's judgements about the protagonist's emotional state in Yuill's study might perhaps be explained in a similarly 'objectivist' way, although here it seems that the children's ascription of a positive or negative valence to an outcome would have to be based on information about what the protagonist wants. On this interpretation, when the children being studied are told that the protagonist wants person A to catch the ball, this marks that outcome as positive, as a good way for things to turn out, while person B's catching the ball remains neutral. The three-year-old child's understanding of the protagonist's emotional responses, on this account, is based on their recognition that people are happy when good things happen and sad when good things don't happen (Yuill, Perner, Pearson, Peerbhoy, & Ende, 1996). The emotional responses at issue can be understood simply as responses to the actual world and how things actually turn out. Crucially, understanding the protagonist's frustration when person B catches the ball does not require the child to track a counterfactual state of affairs to which the protagonist assigns a positive 'valence'—they do not need to be able to think that the protagonist desires that A had caught the ball and would have been happy had that happened.

This 'objectivist' way of thinking, though, is in itself rather limited. The overall valence of a situation, it says, can be altered by whether or not someone wants it to come about. This allows for some idiosyncrasy in that it makes room for the desirability of a state of affairs' depending on what someone wants. It does raise a question, though, about what young children might think that it is for someone to want something. The wanting is something that involves and depends on the person doing the wanting; it is not simply a matter of the state of affairs' being a good one. What is the connection, from the child's point of view, between the information that someone wants such-and-such an outcome, and the value of that outcome?

Wellman and Woolley, in their study on two-year-olds' competence with desire reading, offer an account that might be helpful here. Rather than interpreting children of this age as 'objectivists', Wellman and Woolley suggest that two-year-olds possess a 'simple desire theory' of the mind: they understand desire, not as a representational state with a propositional content, but as a merely intentional or relational state, a state of wanting some

specified object. Here we think not of the desirer's representing some state of affairs as desirable or to-be-brought-about, but merely in terms of the desirer's being attracted to the desired object. Suppose that we do want to say that the infants in Repacholi and Gopnik's study have a somewhat deeper understanding of desire, and that they do have some conception of the experimenter as wanting broccoli. They could nonetheless conceive of this 'wanting' in purely relational terms: broccoli attracts, and crackers repel, grown-ups (or: this grown-up); crackers attract, and broccoli repels, me. Such a relational notion could accommodate a degree of idiosyncrasy without being a matter of the agent's 'point of view' on the world and and without being metarepresentational.

A slight complication is that if we construe 'attraction' as just a matter of the desired object's having a kind of controlling force over the agent, a question remains about why getting the thing that attracts you makes you happy and not getting it makes you sad, and (therefore) how such a relational conception of desire could be connected with the idea of a situation's having a positive or negative 'valence'. This connection might simply be learnt through experience. Children, it might be argued, have plenty of evidence of the connections between getting what attracts you and feeling happy, and between not getting what attracts you and feeling unhappy.

Combining these two lines of approach, then, it seems entirely possible to give a model of under-fours' competence in tests such as those posed by Repacholi and Gopnik, Yuill, and Wellman and Woolley, without crediting those children with any capacity for metarepresentation. In these tests, the idiosyncrasy of the agent's desire can be adequately captured by an 'objectivist' notion of the goodness or badness of outcomes, together with a non-subjective, relational notion of desire. The thinking of children under four may, on this account, be best represented in something like the following way:

- People act to bring about good outcomes and to avoid bad ones.
- People are happy when good things happen and unhappy when bad things happen.
- When something attracts you, it is good to get it and bad to not get it.
- Different things attract different kinds of people. For example broccoli attracts some grown-ups even though it does not attract children. (What kinds of things attract what kinds of people may be to a large extent an open question.)
- People can be attracted by outcomes as well as by objects.

Such generalisations can accommodate a high degree of idiosyncrasy in people's actions and in their emotional responses to outcomes. Nonetheless, grasping these generalisations requires no grasp of genuine subjectivity or a capacity for metarepresentation. On the other hand, the possibility of such an account does not in itself show that under-fours do *not* understand desire as genuinely subjective or as representational. If we take the direct false belief task as the benchmark for metarepresentation, though, and if the idiosyncrasy of desire is to be explained in representational terms, then an account of the sort just sketched would seem to be more parsimonious than holding, on the basis of the discussed findings, that under-fours do after all understand desire in metarepresentational terms. At the same time, if

the idiosyncrasy of desire is not fundamentally representational, such considerations of parsimony might not apply.

4.2.2 Is the relational conception of desire an adequate understanding of its idiosyncrasy?

That under-fours take an objectivist teleological approach to predicting and understanding action, and that they employ a relational conception of desire, are distinct theses and could in principle be employed separately, but they could also be integrated in the manner I have suggested, so that being attracted to something (relational desire) makes getting that thing good for you (teleology). Here the relational conception of desire constitutes a slight modification of, or addition to, the teleological picture. The general picture is that people pursue the good, or what they have reason to do, and are happy when they attain it and unhappy when they fail to attain it. Adding the relational conception of desire to this picture allows us to explain some of the idiosyncrasies in the ways people behave and in the outcomes that make them happy or unhappy. The idea that desire makes a difference to what is good or bad for a person makes room for this additional explanatory power without completely abandoning the teleological schema or simultaneously operating two conflicting explanatory schema. Importantly, it does so without any need for metarepresentation (as would be required if we thought desiring an outcome involved representing it as a state of affairs to be brought about or something one has reason to do). If the only way to make sense of the idiosyncrasy of desire were to treat it as representational, this would obviously not constitute an understanding of desire's idiosyncrasy. However, we might instead think that we have here the beginnings of an alternative model of what an understanding of desire's idiosyncrasy might consist in. If we have seen that ordinary kinds of idiosyncrasy of desire can be accommodated by this model, why think that even a mature understanding of desire requires us to understand it in representational terms?

This line of thought is, as it stands, a little too quick. One reason for this is that a genuine understanding of desire involves an appreciation of desire as something that comes, in some sense, from within the subject. Even if desire can be understood relationally, we would have to understand the relation in question as one that is grounded in something on the subject end, rather than the object end. Compare a relational conception of perceptual experience. We might, as naïve realists do, think of visual experience as being relational in the sense that it is constituted by the perceiver standing in a certain relation to an object, the relation of seeing. Understanding the nature of this relation, though, means appreciating that the 'action', so to speak, is on the side of the perceiving subject: it is the subject who is exercising the power to perceive, and who is modified by standing in the relation of perceiving.

We can contrast this with one possible way of understanding the relational conception of desire. It could, for all we have seen, be that children understand desire quite literally as a sort of attraction. Desired objects, on this model, are like magnets. The idiosyncrasy of desire

comes just from the fact that different objects differentially attract different agents, but the power to attract is in the object, and the desirer is wholly passive. While we sometimes talk about desire in this way even as adults, I take it that such talk is best understood as being largely metaphorical. To understand the way in which desire can be idiosyncratic, we need to appreciate the way in which individuals fix upon or 'choose' certain objects, and not the other way around.

Those who assume that the idiosyncrasy of desire is grounded in its representational nature will hold that this is the sense in which it is grounded in the subject: the subject mentally represents what it is that they desire. But note that this is a proposed *explanation* of something, desire's subjectivity, that we have already characterised in independent terms. The representational idea is that this subjectivity is a matter of the subject's having a picture of the world that can fail to fit how the world actually is. At least in the case of belief, the representation is also objective in a sense—it has objective import, it purports to be about the real world.[99] Understanding the subjectivity of belief is a metarepresentational task because we need to see the agent as mistaking a counterfactual state of affairs for an actual one, and to understand their action in the real world as being based on a falsehood that they take for a truth. This is why it is natural to talk about the subjectivity of belief as the subjectivity of the agent's 'point of view' or 'perspective'—it is a fallible subject's take on how things really are.

While an understanding of desire as representational *could* play a role in explaining the idea that desire can make the desired object good for the desirer and that it is grounded in something in the subject, it is not immediately obvious how it would do so, and nor is it obvious that the subjectivity of desire *has* to be explained in such terms. Recognising this, we should consider whether there are any strong grounds for thinking that a mature understanding of desire must be metarepresentational.

4.3 Conflicting desires

If a mature understanding of desire were metarepresentational in the same manner as a mature understanding of belief, it should be possible to devise a test, analogous to the Sally–Anne style false belief test, that children would only be able to pass once they have acquired that understanding. After all, if coming to understand the true representational nature of desire is a significant development, it ought to make some difference to what we can actually make sense of—to what we can do with the concept of desire. We have seen that while desire's being representational is a possible explanation of its idiosyncrasy, much of that idiosyncrasy can be captured without conceiving of desire as representational. Is there a kind of idiosyncrasy that we can make sense of only with a genuinely metarepresentational concept of desire?

Authors attempting to address this question generally agree that the key thing that a metarepresentational concept enables one to do is to understand cases in which there is

[99]. Compare the objectivity of perceptual experience as discussed for example in (Eilan, 2011).

some kind of *conflict* or *incompatibility* between the content of a given agent's representational mental state and either (i) the content of a state of the same kind belonging to another agent or (ii) the way things actually are. This is clear in the case of belief: a representational conception of belief enables us to coherently represent someone as believing something false, or to represent two people as having mutually incompatible beliefs. As Perner et al. (2005) put it, metarepresentation enables us to resolve 'perspective problems'. In a perspective problem, the contents of two attitudes cannot be conjoined without contradiction, so that integrating those contents into a coherent world-picture requires one to represent them as (merely) represented. Cases of incompatible desires or desires that conflict with reality are thought to be comparable to the cases used in direct false belief tests, wherein what the agent believes cannot, without contradiction, be integrated with reality. This contrasts with tests like the one posed by Repacholi and Gopnik, which require some appreciation of the fact that people can want different things but not that two people can simultaneously want incompatible things. In that study, the infant's and the experimenter's desire-contents are compatible, whether we think of those contents along the lines of 'A eats biscuits' and 'B eats broccoli', or of 'A has reason to eat biscuits' and 'B has reason to eat broccoli'. Those who assume that a mature understanding of desire's idiosyncrasy is fundamentally representational will argue on this basis that none of the studies discussed so far demonstrates that under-fours understand the true idiosyncrasy of desire, since success on the measures employed does not require metarepresentation. Some of these authors have attempted to remedy this situation by devising tests that, according to them, really do test for metarepresentational understanding of desire.

The results of these tests have been somewhat mixed. Some take their findings to support the view that children acquire a metarepresentational conception of desire before they can pass the direct false belief test, while some reach the opposite conclusion. As we will see, though, most of these studies fail to test for genuine metarepresentational ability. For the most part, they do not even succeed in devising tests that would require the ability to metarepresent.

4.3.1 Mixed findings

In an unpublished study, Lichtermann[100] tested children's grasp of incompatible desires by examining their understanding of the emotional responses of two agents to an outcome that ought to frustrate one and satisfy the other. Children were presented with two scenarios. In one, the protagonists' desires are genuinely incompatible, whereas in the other the fact that one agent is satisfied and the other frustrated is merely a coincidence. In one story, for instance, a boy and a girl are travelling down a river that forks. The girl wants to go down the left fork of the river while the boy wants to go down the right fork. In the 'compatible desires' version of the story, each child is travelling in his or her own boat, whereas in the 'incompatible desires' version, they share a boat. In both versions, both the girl and the boy

[100]. Reported in (Perner et al., 2005).

go the same way, but crucially in the compatible desires version, they *could have* gone different ways—that is, their desires could both have been satisfied, and hence are consistent. This is not the case in the incompatible desires version.

Lichtermann tested children's ability to correctly attribute emotions of satisfaction or frustration in the two versions of the story. Children were asked, first, which character was happy, and second, whether the other character was happy or sad. Most children correctly identified the happy character as the one whose desire was in fact satisfied, but most answered the second question correctly only in the 'compatible desires' version of the story. In the version where both characters were in the same boat, children judged both characters happy, even though one of them did not get what they wanted. This, Perner et al. (2005) argue, is because these children lack the metarepresentational concept of desire that is required to integrate the inconsistent desire-contents into a coherent picture.

A similar study was carried out by Moore et al. (1995). In this experiment, three-to-four-year-old children played a game against a toy character, Fat Cat. The goal of the game is to put together a three-piece jigsaw puzzle of a frog, with a body piece, a head piece and an eyes piece. The eyes do not fit onto the body, so players need to get a head piece before they can get the eyes. Players draw cards: a white card means no action is taken, a red card means the player can take a head, and blue means they can take the eyes. Hence both players want a red card first, and after that a blue. The child and Fat Cat take turns to draw cards. What the child does not know is that the order of cards is fixed so that the child will draw a red card before Fat Cat does. Once this has happened and the child has taken a head for their puzzle, they are asked three control questions and two critical questions, the latter being: (1) 'Which colour card does Fat Cat want now?'; (2) 'Which colour card did you want last time?' Moore et al. found that 7/20 children passed both test questions.

The conclusion Moore et al. draw from their findings is that children at this age are preoccupied with their own present desires and so have difficulty thinking about either another agent's desires or their own previous desires where these conflict with their own present desires. Perner et al. (2005) reject this conclusion on the grounds that it would predict, falsely, that children at this age should be hypercompetitive, when in fact they are highly cooperative. Perner et al. suggest that the actual explanation for children's failure in Moore et al.'s task is that because they lack the capacity for metarepresentation, they cannot simultaneously represent their own current desire (that the card be blue) and either Fat Cat's desire or their own previous desire (that the card be red).

The findings of both of these studies have been challenged. Rakoczy et al. (2007) argue that there are methodological problems in both Lichtermann's and Moore et al.'s studies. Moore et al.'s test, they point out, has a 'complex inferential structure: the child has to infer from which piece is missing for each player to which box is the "good" one for each, and finally from there to which color is desirable from her point of view' (Rakoczy et al., 2007, p. 49). Lichtermann's results, on the other hand, could be explained by the structure of the story and the format of the questions: 'it remains unclear', Rakoczy et al. argue, 'what would have happened if the children had been asked first who was sad'; and 'perhaps children

thought the second person was happy as well because she liked to go with the first character together, even though they went to a place different from where she had originally wanted to go' (Rakoczy et al., 2007, p. 50).

In light of these concerns, Rakoczy et al. re-ran versions of both experiments making relevant methodological alterations. Their alterations to Lichtermann's experiment were the following:

i. Instead of explicitly telling children what the two characters wanted, the puppet characters themselves implicitly expressed their desires ('The boat should go to the left/right').
ii. In the case of incompatible desires, the two characters then quarrelled (A: 'The boat should go to the left'; B: 'No, the boat should go to the right').
iii. There were two pairs of questions children were asked after the boat or boats had gone to one side: first, the desire questions as in 'memory for complements' tasks, 'Where did A want that the boat go?' and 'Where did B want that the boat go?'[101] (Q1). Second, the desire-dependent emotion questions 'Is A happy or sad now?' and 'Is B happy or sad now?' (Q2).
iv. In order to accustom children to the questions about desire-dependent emotions of the two characters (Q2), at the beginning of the session a short pre-test was used in which children were asked about the desire-dependent emotions of one single character. This pre-test was included because informal piloting suggested that German children this age often did not read questions about characters' emotions in the required intentional sense (happy/sad *about* something), but rather in an undirected mood sense (happy/sad in general). The pre-test thus presented a baseline for children's proficiency with using 'happy' and 'sad' in intentional ways. Furthermore, children were corrected if necessary, and so the pre-test presented an introductory training to use 'happy' and 'sad' in the intentional rather way for those children who did not yet do it this way.

The authors found, using this modified methodology, that the children they tested did considerably better on both Q1 and Q2 than on false belief tasks, for both compatible and incompatible desires scenarios. They achieved similar results in a test using a similarly simplified alternative to Moore et al.'s card-taking game, and these results were replicated by Rakoczy (2007). The authors conclude that there is a genuine asymmetry in the development of 'subjective' concepts of belief and desire. However, this conclusion rests on the thought that in order to succeed in the tests, children must be operating with a genuinely 'subjective' conception of desire. But it is unclear whether that is really the case, at least if we understand 'subjective' to mean 'metarepresentational', as the authors appear to.

[101]. The study was conducted in German, in which, apparently, this 'want that' construction is more standard than it is in English.

4.3.2 Discussion of these findings

In order to understand what is going on in the story well enough to answer the critical test questions correctly, the children in Lichtermann's study need some way of grasping two thoughts: first, that different outcomes have different valences, and second, that these valences can be different for different people. Even if Lichtermann's original findings and conclusion were correct, the children studied would need to have some kind of appreciation of this; after all, they did correctly identify which character was happy. One way to explain this would be to see the child as identifying with one of the characters in the story; most obviously, when the child is asked who is happy, that might prompt them to identify with the happy character. The child then sees the outcome simply as good, as a happy way for the world to be, and so thinks (at least when prompted) that the other character will be happy too. Perhaps if they had been asked who was sad first, they would have identified with the other character and said that both are sad.

If all a child can do to understand the valence of an outcome for the characters in the story is to think of that outcome as objectively good or bad, then they will not be able to answer correctly all of the questions they are asked about the characters in the story. So if Rakoczy et al.'s results are accurate, and three-year-olds are capable of correctly identifying one participant as happy and the other as sad, then there must be more to these children's grasp of the situation than a mere ascription of objective goodness or badness. One way to complicate things would be to relativise the goodness of the outcome: if the children can think of the outcome as good for the one character and bad for the other, this might enable them to correctly predict who will be happy and who sad. This relativised notion of objective goodness is not itself metarepresentational. 'X is good for A' does not mean that X is good 'from A's point of view', or that A takes X to be good, but simply that X benefits A. This may or may not be a conceptual sophistication that children under a certain age lack, but it is not a metarepresentational notion. Understanding that ericaceous compost is good for camellias but bad for pinks, for example, does not require one to think of camellias as representing anything; indeed, to think that that was what was meant would be precisely to misunderstand the claim.

This makes room for a further twist on the teleological schema and a more sophisticated objectivist teleology. Someone who understands that an outcome can be good for A and bad for B is not restricted to merely labelling outcomes with a positive or negative valence; they can appreciate that a single outcome might have positive valence for one agent and negative valence for another. This does not in itself involve thinking of either agent as mentally representing—in particular, it does not involve representing the frustrated agent as representing a counterfactual state of affairs that would have satisfied her. The application of this schema to the studies discussed above is clear: someone who can understand that the outcome is good for A and bad for B can correctly predict A's and B's emotional reactions without employing metarepresentational concepts.

We should however consider, as we did in discussing Yuill's findings, *why* and *how* children might 'label' the outcome with the relevant (relativised) valences. We, of course, can say that when the boat goes down the left fork this is good for the girl and bad for the boy because she wants it to go to the left and he does not, but can young children understand the case in this way? We come back, then, to the question whether non-metarepresenting children might have any way of understanding the agents' different desires.

On a metarepresentational conception of desire, the relation between desire and goodness or badness might be fleshed out as follows. Getting what you want makes you happy or satisfied and not getting what you want makes you unhappy or frustrated. Whether you have got what you want is a matter of whether the world is as it is represented by your desire. The first child is happy because there is a match between her desire and the world; her desire is fulfilled. The second child is unhappy because his desire is unfulfilled; there is a mismatch between his desire and the world. Things are not as he wants them to be: he wants that the boat went to the right, and the boat did not go to the right. If we want to explain the child's understanding of the case without supposing that she has a capacity for metarepresentation, though, we will not be able to flesh things out in this way. But might there nonetheless be a way to make sense of the case in terms of conflicting desires?

Perhaps there is. Recall the relational conception of desire that I discussed earlier in order to explain children's competence in Yuill et al.'s study. At first blush, such a conception of desire appears to offer little help in accounting for children's success in Rakoczy et al.'s version of Lichtermann's study. The conception in question was that of someone's being attracted to some object, but in the Lichtermann story the thing the children in the boat want is not an object but a certain outcome or state of affairs. However, once we have the relational conception of desire in view, there is no obvious reason why it could not be extended to include objects (in the sense of things wanted) other than objects (in the sense of material things), and there is no obvious reason why what the agent's desire relates them to could not include something more abstract than a material object. After all, if desiring is conceived of as a relation, it would seem to be an abstract, 'intentional' relation (like *seeing* or *thinking of*, for example). If wanting is an abstract relation, though, there is no obvious reason why it should not be able to relate agents to abstracta such as event-types. So, for example, we might say that, in the Lichtermann-style story, one character is attracted to one location while the other character is attracted to a different location; or we could say that one is attracted to the prospect of the boat's going one way, and the other is attracted to the prospect of its going the other way. The actual outcome therefore has a positive valence for one character and a negative valence for the other. In Rakoczy et al.'s version of the study, the child is not simply told what each character wants, but they will be able to assign the right valences to the right characters in light of their quarrelling behaviour. And, recognising that the one character was attracted to this outcome while the other was repelled by it, the child can correctly predict that the first will be happy and the second unhappy.

In the case of false belief, the metarepresentational challenge is to recognise the possibility of mistaking a falsehood for the truth and so misconceiving how things really are.

Understanding the subjectivity of belief requires metarepresentation because in order to fit what the other person thinks into a coherent picture of the world, you need the idea that they have a point of view on the world that can deviate from how things really are. You need to be able to do that in order to predict and explain the other person's actions, feelings and so on, because they will act (feel, etc.) as if the things they think are true: their real and sincere actions are based on a false picture of how things stand. Where desire can be understood on the model of the agent's being drawn towards some object or some possible action or outcome, and of their being happy if they end up getting what they wanted and frustrated if they do not, it is not clear that, insofar as this 'being drawn' is something idiosyncratic, this idiosyncrasy is best thought of through the metaphor of 'point of view', or through the idea that the desirer represents a desired state of affairs as desirable or as something they have reason to bring about. Wanting things to be a way they are not is not, whereas thinking things are a way they are not is, misconceiving reality.

4.3.3 'Conflict' with reality

There might nonetheless be some reason to think that for some desires, at least, we must understand the agent as having in mind a proposition that is incompatible with how things actually are. Harrigan et al. (2018) investigated another kind of 'conflict' in desire, namely a conflict between a desire's content and reality. They tested whether children could reliably judge the truth-value of attributions of 'counterfactual' desires, desires 'about a *concurrent* state of affairs' (Harrigan et al., 2018, p. 4). The authors' thought is that understanding desires that explicitly concern the present time but which are unsatisfied requires an understanding of desire as truly subjective—and here again the authors seem to have a (meta)representational conception of 'subjectivity' in mind—presumably because, they think, it requires the interpreter to be able to compare the represented state of affairs with the actual one and appreciate that the latter fails to fit the former. This kind of test would appear to be, in this respect, the most closely analogous with the classic false belief test, in which the interpreter must appreciate the way in which the protagonist misrepresents the way things presently stand.

Harrigan et al. argue that to test understanding of the possibility of conflict between desire and reality, it is necessary to make the protagonist's desire explicitly present-directed. Ordinarily, they point out, desire-ascriptions are, or at least can be, interpreted as ascribing a desire directed towards the future. In Rakoczy et al.'s study, for instance, children might, as I suggested, be thinking of the protagonists as looking forward to a possible outcome, and Harrigan et al. point out that these children might, even after the story has concluded and one of the characters is unsatisfied, be thinking of that character as still wanting the boat to go to the other side of the lake (at some point in the near future), and hence be ascribing only a future-directed desire to that character. Unless it is certain that the boat will not go to the other side in the future, we cannot be sure that the future-directed desire actually conflicts with reality. Hence we cannot be certain that the child is in fact attributing a

counterfactual desire. Harrigan et al. thus argue that the clearest evidence of understanding 'subjective' desire will be provided by presenting children with scenarios in which an agent wants things to *now* be different from how they in fact are.

To test children's comprehension of such 'counterfactual' desires, the authors presented them with different versions of a story in which a child, Megan, is out shopping with her mother and is sitting in the shopping cart. In one version, Megan's mother ('Mom') asks her to stay in the cart while she gets something from the next aisle. In the other, Mom asks Megan to get out of the cart and fetch some cereal, again while she herself gets something from the next aisle. Megan then either does as she is told or does the opposite. The experimenters had children assign truth-values to the sentences 'Mom wants Megan to be sitting in the cart right now!' or 'Mom wants Megan to be getting cereal right now!' They found that three-year-old children, who would not be expected to pass the direct false belief test, were quite proficient at this task. This, they suggest, shows that such children have a genuinely 'subjective' conception of desire as something that can, like false belief, conflict with reality.

This test assumes that the 'wants … right now' formulation disallows a future-oriented reading of the subject's desire, and that hence if children assign the correct truth-values to desire-attributions of this form, they must understand the desire ascribed as present-directed. We might reasonably be sceptical of these assumptions. Note that 'now', as part of a desire-attribution, can function in at least two distinct ways: it can, as the authors suggest, provide a part of the desire's content, but it can also mark the time as which the agent has the desire in question, without bearing on the content of that desire. 'I want a drink now' is ambiguous between my now wanting a drink and my wanting to have, at the present moment, a drink. It might well be more natural for us to read the sentences used in Harrigan et al.'s study as attributing a desire with a certain temporal index in its content, but given that there is an alternative interpretation of the '… right now' available, the authors are perhaps too quick to assume that children could not be interpreting the sentences as saying that the subject presently has a certain desire, which desire they may think of in future-oriented terms. There is, it seems, nothing to rule out the possibility that three-year-olds interpret 'Mom wants Megan to be sitting in the cart now' as saying that right now, Mom wants Megan to sit in the cart—a desire which could be satisfied by Megan getting back in the cart. Something like this would be enough to explain their answering correctly, but it does not require any understanding of desires as potentially conflicting with reality in the relevant sense.

We might also question another aspect of Harrigan et al.'s methodology. The other studies we have considered tested children's understanding of desire by requiring them to do some simple psychological reasoning, working out what the protagonist of a story could be expected to do or feel given that they have a certain desire. Harrigan et al.'s study, by contrast, only required children to assign the right truth-value to a desire-ascribing statement. Notably, whether the desire-ascription is true or false directly tracks whether the statement echoes what Mom said in the story: when Mom says 'stay in the cart', 'Mom wants

Megan … in the cart …' is true, and 'Mom wants Megan … getting cereal …' is false; when Mom says 'get some cereal', 'Mom wants Megan … getting cereal …' is true, and 'Mom wants Megan … in the cart …' is false. Even if understanding the ascription itself does require a metarepresentational capacity, then, it might be that the children get the truth-value right not because they actually understand exactly what the ascription is saying about Mom, but simply because they remember what Mom had earlier said.

4.4 Two kinds of desire

While I doubt whether Harrigan et al.'s study manages to demonstrate a metarepresentational conception of desire on the part of three-year-olds, the scenario they employ is worth considering further, because I think that it can, in a way that those used by Lichtermann, Rakoczy et al. and Moore et al. cannot, be used to illustrate the existence of a kind of desire that we might want to think of as representational in the propositional attitude sense. However, the scenario can at the same time be used to illustrate another kind of desire that seems not to be representational in that sense. And it turns out that it is the latter kind of desire that is important for the present discussion—it is the kind that is at play in Hampshire's example. The contrast can be brought out by considering differences between Megan's desire and her mother's, and specifically by considering the different ways in which these desires relate to motivation, and to feelings of satisfaction and frustration.[102]

Consider Mom's desire. While I suggested that the children studied might interpret 'Mom wants Megan to be sitting in the cart right now' as ascribing a future-directed desire, perhaps a desire for a certain sort of outcome, it also seems that Harrigan et al. are right to think that there is another interpretation available, on which Mom desires that a certain temporally-indexed proposition be true: Mom would like it to be the case that Megan remains sitting in the shopping cart for the entire period during which she (Mom) is in the next aisle. While I argued that it was not necessary to think of Mom's desire in metarepresentational terms in order to correctly say whether the ascription of that desire was true or false, perhaps we do need to think of it in those terms if we are to properly understand the ways in which such a desire is connected with motivation and with feelings of satisfaction and frustration. However, these contrast with the ways in which Megan's desire, the desire *to get cereal*, is connected with motivation and satisfaction or frustration. Hence, even if Mom's desire is to be properly understood metarepresentationally, the contrast with Megan's desire suggests that the latter desire might not be properly understood in the same terms.

First, Mom's feelings of satisfaction or contentment, or frustration, annoyance or regret, depend on her *epistemic* perspective with respect to how things turn out. Mom will feel satisfied or content just insofar as she believes Megan will stay (is staying, has stayed) in the cart, and annoyed or frustrated insofar as she believes Megan has left or will leave the cart. To the extent that she is uncertain whether Megan will stay or has stayed in the cart, she may feel

[102]. The main observations and arguments of this section are based on (Martin, 1999).

anxiety or apprehension, which will resolve into either satisfaction or frustration when her uncertainty is resolved. Note that *when* Mom's epistemic perspective changes need bear no particular connection to the time of its satisfaction or frustration. She could, confident in Megan's good behaviour, feel satisfied or relieved that she will stay in the cart even as she leaves to go to the next aisle. She might, returning and finding Megan in the cart, still worry that Megan might have at some point got out and then got back in, or she might feel relieved seeing Megan in the cart but then annoyed when Megan tells her that she did get out. If at any point she feels anticipation, it will be when she is heading back to the cart, and what she anticipates is not getting or doing what she wanted, but *finding out* whether her desire was fulfilled.

Now consider Megan's desire to get some cereal. Let's suppose that Megan really is concerned to *get* cereal, rather than simply being concerned that they have cereal or take some cereal home. As children often do, she has got it in her head to do something herself, an action which may be all the more attractive to her for its being an act of defiant mischief. We should here, I want to argue, understand Megan's desire as directed toward a type of act or event. She is concerned not primarily with the truth of a certain proposition; she is concerned *to do* something. One way in which this distinction is salient is in her feelings of anticipation, satisfaction, frustration and so on. For Megan, these depend primarily on her temporal, rather than her epistemic, perspective on the event with which she is concerned. She will feel anticipation up until the point where she acts, excitement as she gets out of the cart, satisfaction when she returns with the cereal or frustration if something prevents her. Even if she is utterly confident of success before she acts, she will not feel in advance the pleasure she feels when she does act—although she may anticipate it.

So, while we arguably need to understand Mom's emotional responses primarily in terms of whether the world is as she would like it to be, this seems not to be the case when we consider Megan's desire. We can understand Megan's desire as directed on a certain type of event—her getting cereal—and as being satisfied when an event of that kind actually happens, and it is the event itself that causes Megan's satisfaction.

The other contrast I want to draw is in the ways in which the two kinds of desire motivate. This second contrast is connected with the first, in particular with the issue of frustration. Megan's desire is capable of motivating action in a way that Mom's is not, and this motivational character gives further reason not to take Megan's desire to be directed fundamentally upon a propositional content.

Desire can only motivate action when its object is seen by the agent as potentially attainable. Assuming a desire with a propositional content must contain a temporal index within that content—which we will if we suppose that propositions have tenseless truth conditions—this means that if, for example, one's propositional desire is for a certain event to occur within a certain time frame, then once that time frame has passed there is no longer any possibility of the desire's being satisfied and hence no possibility for the desire to motivate action. The content of the desire determines the point at which the time for action will have passed. Conversely, if what one desires is that no event of a specified kind occur

within a certain time frame, then the time for action will have passed as soon as such an event occurs.

Suppose that, contrary to Mom's wishes, Megan does get out of the shopping cart whilst Mom is in the next aisle. If what Mom wants really is that Megan stay in the cart for the specified period of time, Megan's action means that Mom's desire is not only unsatisfied, but is unsatisfiable. This desire can no longer motivate action on Mom's part (though a related desire, like a desire for Megan to get back in the cart, might). All Mom's preference for Megan not to have got out of the cart can do is to persist as a lingering disappointment in Megan or regret at her bad behaviour. The matter is settled by the content of her desire's having turned out false.

Things are not so straightforward when we consider Megan's desire. To construe the content of Megan's desire propositionally, we would need to give it an explicitly temporal component. Such a content would tell us in advance when the time for action will have passed. However, if what Megan desires is just to get cereal, this is not something we can specify independently of how things happen to turn out—in particular, it is not something we can specify independently of how long Megan's desire persists. It cannot, for instance, be that Megan desires that she get some cereal at some unspecified point in the future; she would have an entire lifetime to fulfil such a desire, and hence no sense of urgency to act on it in the present. A better candidate for the relevant time period would be: while Mom is in the next aisle. But if what we ascribe to Megan is just a desire to get cereal, we have no particular reason to think that what Megan wants *is* just to get some cereal while Mom is in the next aisle. She might, apprehensive about being caught and told off, sit dithering about whether to leave the cart for the whole time that her mother is in the next aisle. When Mom returns, she may indeed feel disappointed, but there is no reason to assume that this will put an end to the matter: Megan *might* give up at this point, but equally she might anxiously await her next opportunity, determined not to make the same mistake next time. Even when the shopping trip is over, Megan's desire might persist, perhaps even growing in intensity, until the next time they go to the shops, or until she finds a chance to sneak out of the house in a search for cereal—or more likely, but *not necessarily*, until she resigns herself to disappointment and forgets about the matter. Her desire is not determinately frustrated until either she gives up on it or satisfying becomes genuinely impossible.

If at any point Megan does succeed in her mission to get cereal, a particular event will have satisfied her desire, and there is thus a way in which her desire will be satisfied by the way that the world turns out. In this respect, it is like Mom's desire, and we might even posit that there is a set of satisfaction conditions for Megan's desire—a set of propositions such that, if any turn out true, her desire is satisfied. Where Megan's desire differs from Mom's is that this set of propositions is open-ended and cannot be determined in advance of how things turn out. For this reason we should not think of it as the fundamental content of her desire.

4.5 Desire, representation and idiosyncrasy

I have argued that there is a kind of desire that is plausibly thought of as having propositional content, but that there is also another kind of desire—desire for a type of act or event, and also, I would suggest, desire for an object—that is not. I have argued that developmental psychologists attempting to determine when children develop a metarepresentational concept of desire have for the most part failed to devise scenarios which actually require a metarepresentational concept of desire in order to be understood. Regarding the final study discussed, I suggested that while the authors failed to show convincing evidence of metarepresentational desire reasoning in children under four, the scenario they used might nonetheless illustrate a kind of desire that could arguably be best construed in representational terms. I also argued, however, that the very features of the relevant desire that can be used to motivate this thought provide a contrast with another kind of desire, the latter exhibiting a more direct connection to motivation and action.

The Scanlonian cognitivist's proposal about desire, of course, was not the same as the conception of representational desire that the studies discussed here attempt to test for. The latter conception sees desire as representational in the sense that, in desiring, an agent represents a state of affairs that they want to obtain, and reasoning about another's desires involves comparing the state of affairs specified by their desire with how things are in the actual world. The Scanlonian idea was that desiring involves representing the object of one's desire as something that one has reason to pursue or bring about. Understanding another's idiosyncratic desires, on this view, involves seeing how, as it seemed to them, they had reason to do something that they did not in fact have reason to do.

The investigation of the developmental psychologists' conception of representational desire, though, was nonetheless relevant to the question whether the Scanlonian picture is plausible, albeit somewhat indirectly. At the beginning of this chapter, we noted an 'asymmetry' in the development of children's abilities to understand idiosyncratic desire and to understand idiosyncratic belief, with the former apparently developing much earlier. If we (provisionally, at least) accept that children become capable of metarepresentation at around their fourth birthday, as I argued that we should, this suggests that understanding the idiosyncrasy of desire does not require metarepresentation—hence the idiosyncrasy of desire is not fundamentally to be explained in terms of desire's representational nature. In this, desire contrasts strongly with belief. And this reinforces the claim, already motivated in the previous chapter, that the idiosyncrasy of desire-attributing rationalisations is of a different kind than that involved in perspectival rationalisations. If subsequent studies had either turned out to find good evidence for a metarepresentational understanding of desire in under-fours, *or* to successfully illustrate that under-fours fail to appreciate something essential to the nature of desire's idiosyncrasy, this line of thought might have been undermined. I argued that the studies in question failed to achieve either of these goals, at least if we keep our focus on the kind of desire that is at issue in the present investigation: desire for an object or desire to act in a certain way.

One loose end remains to be tied up. I suggested that children under four might have an understanding of desire as attraction to an object, and that they might appreciate the idiosyncrasy of desire in those terms. We saw, though, that this raises a question as to whether those children really understand the *subjectivity* of desire—the way in which its idiosyncrasy is grounded in the subject rather than the object.

While it is hard to say anything definitive on this issue, we have seen no real reason to think that these children *must* be understanding desire's idiosyncrasy as based entirely in its object. Moreover, I think there is at least some reason to think that even young children do have some appreciation of the subjectivity of desire. A conception of desire as coming wholly from the object sees desire as a kind of magnetic force, something that comes from outside oneself and controls one's actions. It does not appear that young children feel themselves beset by their love of biscuits or their dislike for broccoli in this sort of way. It seems reasonable to say that, in some sense, a child will typically 'identify' with his or her desire for biscuits, if the alternative to identifying with a desire is experiencing it as alien or as coming from outside oneself.

Another apparent advantage of cognitivism about rationalisation, though, is that it provides an explanation of the nature of this 'identification'. The idea is that one identifies with one's desires insofar as one can take oneself to desire what one desires for a good reason, and one feels alienated from one's desires when one can see no reason to desire what one desires. If we want to hold, as I am suggesting we should, that the idiosyncrasy of desire is not the idiosyncrasy of apparent reasons, and that a person can intelligibly desire something that, as it seems to them, they have no reason to desire, we will need to give some alternative account of this pair of notions, identification and alienation. This is the issue I will try to address in the next chapter.

Chapter 5

Being Unalienated

5.1 When do desires make sense to the desirer?

I suggested at the end of the last chapter that children under four might, in operating with a simple relational conception of desire, have what is in effect a perfectly correct and adequate appreciation of desire's idiosyncrasy. Part of that conception, as I said, must be that these children have some grasp of these desires as coming from within themselves; a desire for some object is not an external force that pushes one around, or a power of the object to attract one, but a matter of one's own natural affinity for that object.

Many authors have drawn a distinction, closely related to this point, between desires with which the subject of the desire identifies and desires from which the subject is alienated. While many of our desires are experienced as being truly our own, or as coming from within us, there are some, it is said, that strike us as alien, external, other. In the latter cases, we feel we are in the grip of something that we cannot control, and that is not truly ours. If someone acts on such a desire, that desire is not apt to rationalise the person's action. From the agent's perspective, such an action does not make sense; it is explained not in terms of what seemed from the agent's perspective to be good reasons, but in terms of the agent's failing to resist a powerful motivational force. This raises one of a set of challenges to our account of Hampshire's collector, and more broadly to the view that desires can generate reasons for action, that press on us the need to explain something about the collector's perspective on his own desire. A related challenge comes from Anscombe's argument that what is desired must be desired under the aspect of some desirability characterisation. If that argument is compelling, it puts pressure on us to say in what way the collector sees the inferior bronze as desirable. But, we might think, any desirability characterisation of the bronze will correspond to a worldly reason for buying it, thus calling into question our claim that it is his desire and his desire alone that is his reason for acting.

The challenge to explain the subject's identification with their desire and the challenge to provide a desirability characterisation seem to be related, at least in that some have

appealed to something like the Anscombean idea in order to explain what is involved in a subject's identifying with a desire. Identification, these authors argue, involves seeing the desired object as being good or desirable. I will argue that the two challenges are in fact importantly different. Anscombe's argument needs to be dealt with separately from the concern about alienation, because a desire can meet Anscombe's condition whilst still being one from which its subject feels alienated.

Dealing with the concern about alienation first, I will argue that the demand for a positive account of identification is in fact misplaced: it is plausible that we understand *alienation* in positive terms and that identification is simply the default case, the relation we bear to our desires when not alienated from them. There is therefore no need to say more about why Hampshire's collector does not experience his desire for the bronze as alien. However, there is still the need to say something about the respect in which he sees buying the bronze as desirable. My answer to this will be that a desirability characterisation can be provided by the desire itself.

5.1.1 Some deviant examples

In Chapter 3, I briefly commented on an appeal that Scanlon makes to a famous example from Warren Quinn, the example of the 'radio man'. Here is that example as Quinn himself presents it:

> Suppose I am in a strange functional state that disposes me to turn on radios that I see to be turned off. Given the perception that a radio in my vicinity is off, I try, all other things being equal, to get it turned on. ... [T]his is all there is to the state. I do not turn the radios on in order to hear music or get news. It is not that I have an inordinate appetite for entertainment or information. Indeed, I do not turn on the radio in order to *hear* anything. (Quinn, 1994, p. 236)

Insofar as we can think of the radio man as *desiring* at all, he seems to be alienated from his desire to turn on radios. His desire does not make sense to him, and neither do the actions he takes because of that desire. There is no rationalisation here: there is no story that reveals what point there is, from the radio man's perspective, in his turning on radios. There is no such point. There seems to be something missing from the radio man's psychological situation. He is not, as we might say, identified with his motivational state; he is merely passive with respect to it.

It is somewhat doubtful whether we should really think of the radio man's disposition to turn on radios as a desire at all. However, the idea of alienation from one's desires can be illustrated by more realistic examples in which something more like an actual desire does play a role. Consider three of Gary Watson's: a mother feels the urge to drown her screaming child in the bath and is horrified with herself; a humiliated squash player feels compelled to smash his racquet into his opponent's face, even though he is disgusted by the thought of behaving in such a bestial way; a very religious man detests his sexual urges, which he believes to be the work of the devil.[103] The urges in these cases are perhaps more recognisable as

[103]. (Watson, 1975).

something on the order of desires, but still they do not stand to rationalise action in the way that the collector's desire rationalises his buying the trashy bronze. If the agent in any of these examples acted on their urge, their action would not make sense to them in the way that the collector's makes sense to him. Perhaps there would be some kind of 'inside story' to tell, but it would not be one that revealed the action to be, from the agent's perspective, rational or justified.

We have, then, a division within motivation: on one hand, there are desires with which the subject identifies, which seem (potentially at least) fit to rationalise action; on the other, there are mere urges, motivational states the subject experiences as alien forces to overcome or be overcome by. This presents a double challenge to the claim that desires can provide reasons for action. First, the proponent of that claim must provide some account of the division and hence of how some desires can provide reasons while others do not. Second, they must do so without making something other than desire, in the case of identification, do the real rationalising or reason-giving work.

The latter part of the challenge arises because one attractive explanation of what is missing in these examples of alienated motivation is the agent's seeing a good reason to do what they are motivated to do. The problem with the radio man is that he cannot say *why* he wants to do what he wants to do, where being able to say why, here, would be a matter of seeing some (worldly) reason to do it. Or perhaps, given that the radio man might recognise possible reasons for turning radios on that are not *his* reason, we should say more specifically that identification with his 'desire' would require him to see some reason to turn radios on that was from his point of view his reason for wanting to turn radios on—the kind of thing he would cite if asked why he wants to turn radios on.

The idea that there is a connection between intelligibly wanting and having an answer as to *why* one wants what one wants is put forcefully by Anscombe. In §§37–40 of *Intention*, she argues for a certain 'relative' restriction on 'possible objects of wanting'. As she puts the thought in the analytical table of contents at the start of the book, the restriction is that

> If a man wants something, he can always be asked what for, or in what respect it is desirable; until he gives a desirability-characterisation. (Anscombe, 1963, p. viii)

There are two parts to this idea: (i) that someone who wants something can always be asked what for or in what respect it is desirable; (ii) that repeated application of this line of questioning will always terminate with a 'desirability characterisation' (or else we will not understand the claim that the person wants what it is claimed they want). Anscombe illustrates the point in typically memorable fashion:

> But is not anything wantable, or at least any perhaps attainable thing? It will be instructive to anyone who thinks this to approach someone and say: 'I want a saucer of mud' or 'I want a twig of mountain ash'. He is likely to be asked what for; to which let him reply that he does not want it for anything, he just wants it. It is likely that the other will then perceive that a philosophical example is all that is in question, and will pursue the matter no further; but supposing that he did not realise this, and yet did not dismiss our man as a dull babbling loon, would he not try to find out in what aspect

> the object desired is desirable? Does it serve as a symbol? Is there something delightful about it? Does the man want to have something to call his own, and no more? Now if the reply is: 'Philosophers have taught that anything can be an object of desire; so there can be no need for me to characterise these objects as somehow desirable; it merely so happens that I want them', then this is fair nonsense. (Anscombe, 1963, pp. 70–1)

One kind of answer to the 'Why?'-question, of course, is instrumental: one wants *x* because having *x* will help one to attain *y*, which is something else one wants. This kind of answer, however, merely raises the question again, unless it is obvious in what respect *y* is desirable. Saying I want a saucer of mud because I want to eat some mud does not really explain my desire for a saucer of mud unless you can see why I would want to eat mud—although it might point to where a real explanation is to be found. If, on the other hand, I want to throw the saucer of mud at my rival, my desire is no longer unintelligible in the same way. You might think such aggression vicious or unjustified, but unless you are so constitutionally pure that you cannot understand the motive at all, you will at least see what it is that appeals to me about having a saucer of mud. Anscombe concludes that the kinds of answer to the 'Why?' and 'What for?' questions that leave 'no room' for further questions of the same kind are those that give a characterisation of the thing wanted as being in some respect desirable, or good: a 'desirability characterisation'. The trouble for the proponent of desire-based reasons is that a desirability characterisation looks a lot like an apparent worldly reason. If simply appealing to further desires cannot make one intelligible as wanting something, it seems as though in the end one has to be motivated by the kind of thing that would be cited in a worldly or perspectival rationalisation. This puts pressure on the idea that desires play a fundamental role in rationalisation. Why not just think that it is after all the agent's perspective on universal reasons that does the real work? On a cognitivist account, a desire one sees no reason for is inevitably experienced as an alien urge. We are active with respect to those aspects of our lives in which we exercise Reason, and passive with respect to everything else.[104]

5.1.2 The question 'Why?'

I have grouped together Quinn, Watson and Anscombe as raising concerns about the first-personal intelligibility of desire, and I have connected this with the idea of identifying with or being alienated from a desire. In fact, the arguments of Quinn, Anscombe and Watson are importantly different, and they raise somewhat different concerns. In a way it is only Watson's cases that directly concern alienation and identification from one's desires, as we will see. The kind of concern raised by Quinn's and Anscombe's arguments also need to be addressed, but, as I will explain, they need to be addressed in different ways.

Quinn's primary concern is to reject a view he calls 'subjectivism', which combines, first, a rather crude functionalism about desire as a simple, 'brute' disposition to act in certain ways, and second, the view that actions are fundamentally rationalised by an agent's desires. The radio man example is intended to show what is wrong with this picture. Quinn argues

[104]. Compare (Raz, 1997).

that what is needed to rationalise his actions is 'an evaluation of the desired object as good' (Quinn, 1994, p. 247), and suggests that this is ordinarily present in desire. However, it is in principle open to the 'subjectivist'—or, indeed, the proponent of the view that *some* actions are fundamentally rationalised by desires—to argue that what is needed is simply a richer or 'thicker' conception of desire, in particular a conception that says something about how it is to desire something from the desirer's perspective.

This point is pressed by David Copp and David Sobel (Copp & Sobel, 2002), who suggest that desires, unlike the radio man's brute disposition to action, are integrated in certain ways with the subject's other mental states, in particular her other motivational states. Copp and Sobel suggest that desiring to do something typically involves at least the tendency to think about doing it, to plan ways to do it, and to object when obstacles are put in the way of one's doing it. Sabine Döring and Bahadir Eker suggest, similarly, that

> the desiderative dispositional profile necessarily includes, roughly, dispositions to form long-term intentions to achieve the object of the desire, to integrate such intentions into more general and complex plans the agent already has, and to form agential policies that encode general patterns of action in certain specific situations. (Döring & Eker, 2017, p. 102)

Merely connecting the desire with further goal-directed attitudes, however, fails to address a worry that is, if not exactly Quinn's, certainly in the close vicinity. This is perhaps best brought out by looking again at Anscombe's argument. Whenever someone wants something, Anscombe observes, we can ask why, or what they want it for. One kind of answer to this question is instrumental, relating the object to a further want—or, indeed, a further intention, plan or policy. Suppose I am looking for a pen. You ask me what I want a pen for and I tell you that I want to draw a picture of a spider. While this reply in a way answers your question, it just raises the same question again with respect to this more general want: Why do I want to draw a picture of a spider? Perhaps I explain that I have a friend who is afraid of spiders and I want to see if my drawing of a spider will frighten him. Once again, you can ask why I want to do that. Anscombe's claim is that such a line of questioning must terminate with an answer that provides a characterisation of the object of my want as desirable in some respect. For example, I might say that it will be amusing to frighten my friend, or that he frightened me and I want to even the score, or that I want to use his reaction as a test of my spider-drawing skills. Answers such as these are apt to make my looking for the pen intelligible as an intentional action, because they show that it has a *point*, and they make me intelligible when I claim to want a pen because they show what I am after in wanting a pen.

Why must the series of 'Why?' questions terminate in a desirability characterisation? Anscombe's argument, which begins in the passage I quoted above, is easily misunderstood. It might seem that what Anscombe is claiming is that a saucer of mud, for example, is just not the kind of thing that can be wanted, because it is not good. To think this would be to miss the point of saying that it is a *relative* restriction that is being placed on objects of wanting. Anscombe is not here proposing a division amongst possible objects of wanting.

She does that in the preceding sections of *Intention*, in which she argues that one cannot want things in the past or things that one takes to be impossible, and the 'desirability characterisation' idea is explicitly introduced as a relative restriction precisely in contrast to such absolute restrictions. The point is not about what kinds of things can be wanted but about the structure of an agent's ends and how the agent understands those ends. Any end a person has, Anscombe is saying, is an end for them under the aspect of some desirability characterisation. Notably, there are not any obvious restrictions on which ends can be characterised as desirable. The significance of the saucer of mud example is simply that—unlike with many things we might more ordinarily want, like food, fun, company, comfort, obviously useful or pleasant objects, and so on—it is not at all *obvious* why one might want it. Given the right context, even Anscombe's examples might have an obvious point. If you know that I am creating a rustic table decoration, or that I am a great admirer of *Intention* and something of a romantic, you might not need to ask what I want a twig of mountain ash for.[105] This does nothing to vitiate Anscombe's point, because what such a context makes obvious is simply a possible characterisation of the object as, from my point of view, desirable.

Anscombe's examples are effective because with many ordinary wants, their objects' desirability being obvious, we are apt to overlook the significance of those objects' desirability in making our wanting them intelligible. In discussing Quinn's example and Scanlon's use of it, Copp and Sobel observe that there need not be anything bizarre about having basic, unmotivated desires, and give as examples the desires to be healthy, to be clean, and to avoid silence (Copp & Sobel, 2002, p. 259). Such basic desires are indeed perfectly intelligible, but they are also desires for things that are, in perfectly obvious ways, desirable. To cite such objects of desire to support the claim that there need not be anything bizarre about having basic desires does not speak to the claim that desires are only intelligible where the desirer sees the desired object as desirable.

5.1.3 Desirability characterisations

So the case for thinking that the desirer must see the desired object as in some respect desirable continues to stand, at least for now. However, there is more to be said about what role exactly these desirability characterisations play in rationalisation.

As well as challenging the necessity of the desirer's taking the desired object to be good, Copp and Sobel also argue that this is not sufficient for making actions intelligible:

> [I]magine that we merely add to the radio man's psychology a tendency to see something desirable in turning on radios. The radio man keeps turning on radios, and finds himself having the thought, 'Wouldn't it be nice if all the radios were turned on now.' This sounds to us more like an obsessive thought process than a desire. (Copp & Sobel, 2002, p. 262)

[105]. A true story: a friend and former colleague of my wife's, with whom I had discussed my doctoral research some time previous, gave me a twig of mountain ash, or rowan, as a wedding present. It remains a treasured possession and is displayed, dried, in a cabinet in our kitchen.

Döring and Eker make a similar point, here concerned with evaluative belief rather than a thought's occurring to the radio man:

> Now, let us suppose that Radioman has the belief that turning on all the radios in his vicinity is intrinsically good. Our question is: Is there any sense in which Radioman's action is even slightly more intelligible or less bizarre now that we imagine him as someone who thinks that turning radios on is a worthwhile activity in itself? We think not! Despite having ascribed to him the evaluative belief in question, we are still puzzled as to why he acts as he does; in fact, now that we assume him to be committed to the idea that turning radios on is intrinsically valuable, the case is even more perplexing, if anything. (Döring & Eker, 2017, p. 95)

There are a couple of ways of responding to this argument, one more concessive than the other.

The least concessive response is that the radio man's thinking that turning on radios is good in itself *is* in principle enough to make his desire intelligible, it is just that it is very hard to see why he would, or indeed how he could, believe this—so it just raises a further but, crucially, distinct interpretative challenge. I do not find this response entirely satisfactory, and the reason why can be put in Anscombean terms: simply by telling us that he thinks turning on radios is *good*, the radio man does not show us what *point* there is in his turning on radios. We want to know: In what way good? Certainly, as Michael Stocker (2011) argues, explicitly evaluative beliefs can make desires and actions intelligible, as when I say (in Stocker's examples) 'I want to get you something good' or 'I want to do what is best'. However, I am inclined to think that such explanations only make for intelligibility because context provides a more substantial content to 'good': I want to get you a good present; I want to do what is morally best, or best in terms of the choice values implicitly understood to be relevant in the situation.[106]

I favour a somewhat more concessive response. The critics are right that simply adding the belief that turning on radios is good does not make the radio man's actions intelligible. However, this is just because such a belief does not show us the point he sees in turning radios on: it does not enable us to see *how* he views turning radios on as desirable. To say 'this is good' or 'this is desirable' is not to give a desirability characterisation; a desirability characterisation must say in what respect, under what aspect, the thing is taken to be desirable.

This is a point already emphasised by Anscombe. She raises, in this context, the Thomistic–Aristotelian idea that the forms of goodness or desirability that can provide suitable stopping-points for the 'Why?' series will fall under one of three heads: 'should', 'suits', or 'pleasant'.[107] Whether or not this threefold distinction is in fact exhaustive, it is a central aspect of Anscombe's 'desirability characterisation' idea that there are many forms of the good: '*bonum est multiplex*' (Anscombe, 1963, p. 75). To understand someone's aim, their wanting, we must have some sense of *which* form of the good they see the object of their wanting as falling under. Notably, Anscombe also claims that while 'the notion of "good"

[106]. For very helpful discussion of the idea of choice values, see (Chang, 1997b) and the introduction in (Chang, 1997c).

[107]. See (Vogler, 2002) for an in-depth treatent of this idea.

that has to be introduced in an account of wanting is not that of what is really good but of what the agent conceives to be good', (Anscombe, 1963, p. 76) nonetheless 'the good (perhaps falsely) conceived by the agent to characterise the thing must *really* be one of the many forms of good' (Anscombe, 1963, pp. 76–7). This is why supposing that the radio man takes turning on radios to be a form of the good gets us nowhere. He could say 'I think turning radios on is good in itself', but this is just a form of words. To begin to fathom what he might mean by it, we would need further characterisation of *how* he sees this as good, and that characterisation would have to connect turning on radios with something we recognise as a genuine form of the good.

All this might seem to suggest that Anscombe is a firm cognitivist. She seems to be saying that when one intelligibly wants something, one's wanting is based on a belief that the thing is good in some specific respect. What kinds of things are good is an objective matter, and we can only understand an action insofar as we can see it as being taken in pursuit of genuine goods. These goods are qualities that the agent takes the object of their want to possess, and it is the agent's taking the object to possess some form of goodness that makes their pursuing it intelligible. I think this is to read too much into Anscombe's argument, and I think we can accept her argument whilst rejecting Scanlon-style cognitivism. I will explain how later on. First we need to discuss Watson's cases and the challenge to explain what it is to identify with a desire. In doing so, we will also clarify something about Anscombe's conclusion.

5.2 The challenge of alienation

It what we might call *Watson-style cases*, an agent feels a motivation of a kind that we can recognise as at least something like a genuine desire. The agent is genuinely attracted or drawn to a certain action but at the same time, for some reason or other, rejects it. Along with Watson's angry tennis player, pious would-be debauchee, and mother at her wit's end, we can recognise Harry Frankfurt's (1971) famous unwilling addict as an example of this kind.

The first thing to note about such cases is that, as I said, they do seem to involve motivational states that are intelligible as something like genuine desires. In this respect they contrast with the example of the radio man. The difference seems to be precisely what our discussion of Quinn's and Anscombe's arguments would suggest: in each case, we can see a characterisation of the object of the agent's urge as in some respect desirable. Smashing his opponent in the face would, for the tennis player, feel like revenge; the pious man is drawn by the lure of sexual pleasure; the mother would, by killing the child, make the screaming stop; the addict would experience the pleasure, or at least the relief, of getting high.

This highlights something important about Anscombe's claim about desirability characterisations. Her claim is that someone is intelligible to us as wanting something only if we can see them as wanting it under the aspect of a desirability characterisation: only when we can see how the desired object is seen as desirable can we understand the person as really

wanting the thing at all. This is a relatively formal claim about desire-ascription. It relates to our understanding of action in a similarly formal way: we can understand a person as acting with a certain aim or intention only if we can see a desirability characterisation of that aim. What Anscombe's thesis does not say is that seeing something as in some respect desirable makes it intelligible *for* one to act in pursuit of that thing, if this means having an adequate rationalisation for doing so. In each of the examples, we can see, in a minimal sense, what the point would be for the agent in acting on their desire, even if their doing so would be harmful, vicious, bestial or even monstrous.

Monstrous actions do pose a genuine challenge to our understanding. The actions of serial killers, for example, may not make transparent sense to us, and getting close to something like an appreciation of how they understood their own actions can be a daunting hermeneutical task. Identifying a 'desirability characterisation' that they took to characterise their actions is only a first, relatively basic, part of this; if Anscombe is right, it is merely a precondition for seeing them as acting intentionally at all. Serial killers, for instance, often seem to kill for some kind of pleasure or gratification, or to 'get their own back' against a group of people they feel they have been wronged by. If we can understand this—and whether we really can is perhaps not obvious—we can find their actions intelligible *as intentional actions*, even if there is still a very good sense in which we find these actions incomprehensible. Much of the difficulty we face in understanding such actions seems to go hand in hand with a difficulty understanding the agent themselves. Understanding a person in this way might require rich historical and psychological investigation and interpretation, and perhaps a good deal of imagination and empathy, that goes much deeper than the idea of 'practical reason'.

Turning our attention back to more ordinary agents, the point about the role of desirability characterisations highlights that one can be alienated from a desire that meets Anscombe's relative restriction: you can see the thing you want as being in some respect good whilst nonetheless feeling that this desire is not truly yours, or that you do not understand it, or that you wish to disown it. Hence we need to say more to explain what has gone wrong in Watson-style cases. A simple cognitivist account might say that an agent identifies with a desire of theirs just in case there is, from the agent's perspective, good enough reason to do what it is that they want to do. This is too strong: clearly we can and often do have desires from which we are not alienated but where we think we have better reason to do something else. Consider an ordinary dieter's desire for a piece of cake, or the desire to clock off and have a glass of wine when you have important work to do. I will not consider how the cognitivist account of identification might be improved; what is at issue in the present chapter is whether we can explain identification *without* appeal to cognitivism and hence understand desire as a genuine source of reasons.

5.2.1 Non-cognitivist accounts

The possibility of alienation from a desire illustrates that our desires only stand to rationalise our actions if we relate to them in the right kind of way. This might seem to suggest that something needs to be added to a mere desire for it to be potentially rationalising—something needs to be added that, as it were, *backs up* the desire, thereby giving it the 'authority' of a genuine reason for acting. The cognitivist can attempt to answer this explanatory challenge by appealing to something external to the agent's subjective motivational states, namely their perspective on the worldly reasons that apply to them. If we want to give an account of how a desire can stand to rationalise action without being backed up by apparent worldly reasons, though, we will need to look elsewhere.

We have already described the contrary of a agent's being alienated from a desire as their 'identifying' with it. One way to go is to take this language quite seriously and to think of identifying as itself some kind of positive mental action, stance or attitude, by which an agent adopts a desire as truly their own. If this is not to be understood in terms of, for example, judging that the desire is adequately supported by worldly reasons, the most obvious alternative is to think of identification as consisting in some kind of higher-order conative state—most straightforwardly, a higher-order desire, such as the desire to have a certain desire, or the desire that a certain desire move one to action.

Watson, criticising a view along these lines that he finds in Frankfurt's 'Freedom of the Will and the Concept of a Person',[108] observes that if we simply appeal to higher-order desires, the question of whether the agent is identified with a given desire can simply be raised again at this higher level. The only ostensible way in which a second-order desire differs from a first-order desire is in its object: a second-order desire is just a desire that concerns a first-order desire. If the issue of whether an agent is 'identified with' some desire can arise at the first order, surely it could also arise at the second. Clearly, simply ascending to yet higher orders of desire won't help. Much as, as we saw in our discussion of Anscombe's argument, simply appealing to further and further ends to which a given desire is subsidiary does not make the latter desire intelligible as such, so simply appealing to higher and higher orders seems not to be able to settle the question whether I 'identify' with my desire.[109]

For present purposes, though, the regress worry is not the most relevant challenge to the higher-order desires proposal. The problem for our purposes is that, once again, it is not clear that it can make sense of the whimsical desirer's perspective on his desire and his action. If we try to understand Hampshire's collector in terms of the higher-order desires approach, we seem to run into the same kinds of problems that we met in trying to give a Scanlonian cognitivist account of his attitude. Suppose we say that he wants to want the inferior statue, and that this is why his desire gives him a reason to buy it. This seems very close to the

[108]. (Frankfurt, 1971). It is not clear that Frankfurt actually endorses anything as simplistic as the claim that identifying with a desire always consists in having a certain kind of higher-order desire, and his view is explicitly subtler than this as developed in later papers. See (Frankfurt, 1988, Chapters 5, 12). Less subtle use of higher-order desires is made by (Lewis, 1989), who analyses the attitude of *valuing* as desiring to desire. (Scheffler, 2010, Chapter 1) provides forceful criticism of this theory.

[109]. (Watson, 1975).

situation of the collector who is ashamed of his bourgeois tastes and wants to be a bit more trashy and subversive. Or, if it is not, we need some account of the difference. If the collector's desire only gives him a reason to buy the bronze *because* he wants to have that desire, it is hard to see how we can make sense of the first-order desire, the desire for the bronze, as being the real source of the collector's reason. We cannot capture the *particularity* of the collector's reason.

There are alternative ways for the non-cognitivist to try to explain the identification-relation. Not all of them raise the kind of regress worry that Watson presses against the higher-order desire account. An attractive common strategy, that might be developed in a number of different ways, will be to attempt to explain the integration of the desire into the agent's psyche in terms of something more general under which it is subsumed—much as, on the cognitivist account, the non-alien desire is subsumed under some universal value, or some general principle concerning what kinds of facts are reasons for what kinds of responses. We might, for instance, propose an expressivist account of valuing as being stably disposed to conduct one's practical life in a particular way,[110] and hold that we identify with desires that manifest our values. Or we might suggest that the desires with which we identify are those that cohere with our other desires,[111] or that the desires that are fit to rationalise action are those that arise in the right kind of way from past experience.[112] However, it seems that insofar as these accounts propose to explain the authority of the desire itself by backing it up with something more general, they will lead us into the same kind of puzzlement when we consider a desire like that of Hampshire's collector. The collector's desire for the bronze *does*, in a way, intrude upon him; it does come unbidden, and it need not cohere or integrate in any straightforward way with his other desires and values. At least, its doing so does not seem to be the source of its significance for him. Indeed, we can perfectly well imagine a situation wherein the desire does, at least at first, conflict with some of his other desires and attitudes, and in which he accommodates the latter to the former because of the original significance it has for him. In fact this seems to me a quite familiar and important process—a kind of self-discovery that most of us undergo every now and then, perhaps most notably and intensely during adolescence, but also later in life. If we only let ourselves 'identify' with those desires which we could readily subsume under something more general, an important source of personal growth and enrichment would be closed off to us.[113]

5.2.2 A deflationary account

I want to propose that the challenge from identification rests on a mistake. We considered examples in which an agent feels alienated from a desire or desire-like state and in which the state from which they feel alienated fails to give them a reason to pursue its object, or at least

[110]. (Blackburn, 1998).

[111]. Perhaps appealing to an account of practical coherence along the lines of (Millgram & Thagard, 1996).

[112]. (Millgram, 1997)

[113]. Peter Railton, in recent work, has begun to develop the somewhat similar idea that desire plays an important role in the discovery of values. See (Railton, 2012) and his 2018 Locke Lectures (available at https://www.philosophy.ox.ac.uk/john-locke-lectures).

fails to give them a reason of the right kind. In light of these cases, we recognised a distinction between desires with which the subject identifies and those from which she is alienated. All of the accounts we have considered so far, cognitivist and non-cognitivist, attempt to give a positive account of identification, implicitly understand alienation as simply the relation we bear to our desires when we fail to be identified with them. This invests the resultant theories of identification with an assumption which on closer inspection might well seem odd, namely that our default relation to our own desires is that of alienation, and only when something extra is added do we see them as truly our own. The examples of alienation give us no reason to accept this assumption.

Absent that assumption, an alternative kind of account is possible. If in each case of an agent's being alienated from some desire of hers there is a positive explanation of that alienation, then 'identification' can be understood as the default: to identify with a desire of yours is simply not to be alienated from it. To identify with a desire, on this view, is something like accepting it, where acceptance can consist in a mere lack of opposition.

If we return to consider the examples of alienated desire, we can see that it is very natural to understand them in this way. In each case, the agent experiences a desire which, for one reason or another, they reject, resist or disavow—because acting on it would be monstrous, ruinous, vicious or bestial; and perhaps, in the case of the addict, because the desire itself is unnatural, not authentically the agent's own but the product of a malign chemical manipulation. The possible reasons for rejecting, or for feeling alienated from, a desire can concern either the desire itself or its object. It is natural to think that when someone rejects a desire for a reason concerning its object, the reason in question will be one that, from the subject's point of view, shows the object of the desire to be bad in some respect: it would be infanticide, for instance, or a shameful act of animalistic aggression. There appears to be as much flexibility here as there is in the requirement that what is desired under the aspect of a desirability characterisation. Anscombe suggests that someone's saying 'Evil, be thou my good' need not be senseless: 'What is the good of its being bad?' could be answered by a 'condemnation of good as impotent, slavish, and inglorious', so that 'the good of making evil my good is my intact liberty in the unsubmissiveness of my will' (Anscombe, 1963, p. 75). For Satan, perhaps a desire to do something good could be experienced as alien. Whether a desire's object's being bad in some respect will lead its subject to feel alienated from it would seem to depend to a great extent on their general mindset, what is important to them, what they cannot stand, and so on.

Considerations of badness could also come into reasons for rejecting a desire that concern the desire itself. Someone might, for instance, recognise that money is good in various ways but firmly believe that only corrupt people desire to become rich. If such a person found themselves craving wealth, they might experience this desire as alien, as not being truly theirs. Perhaps the first class of considerations, concerning the badness of what is desired, might actually be subsumed into this category: the agent does not like to think of themselves as the kind of person who would have such an evil desire.

However, if we look at the desire itself and what it might say about the subject, it is not only explicitly evaluative considerations that might intelligibly lead to the subject's feeling alienated from their desire. Your desires say something about what kind of a person you are, and one kind of reason you might feel alienated from a desire is just that it clashes uncomfortably with your conception of yourself. Perhaps Hampshire's collector would have felt this way if he had strongly identified as someone with purely exquisite taste, rather than being prepared to accept his attraction to the bronze as expressing something that is, so to speak, true to who he is. This sense of a desire's not reflecting one's true self is starkest when the agent believes their desire to have been instilled in them by manipulation, whether this be relatively mundane chemical or psychological manipulation or the kinds of science-fiction scenarios one finds in philosophy papers and accounts of people suffering from psychotic delusions. On the other hand, it could be as mundane as the person's feeling that a given desire is 'out of character'.

Some of the factors that can contribute to a sense of alienation look like worldly reasons. However, this does not undermine the claim that the reasons generated by an agent's desires are not themselves worldly reasons and that their force as reasons cannot be explained by worldly reasons. It may be that the rejection of a desire is something that can be given a worldly or perspectival rationalisation, so that understanding why some desires do *not* give their agents reason to pursue their objects is sometimes something we understand by appreciating the agent's perspective on their worldly reasons. It is perfectly consistent with this that in the normal case, where one is not alienated from one's desire, it is the desire itself, and nothing else, that gives the desirer a reason to pursue its object.

If we consider examples of a person's feeling alienated from a desire of theirs, then, we can see that it is in fact very plausible that alienation is a matter of the agent's taking some kind of active stance against the desire. A diverse range of factors can lead to this situation, but in each case there seems to be some positive explanation of the agent's alienation: the explanation is not simply the absence of some extra condition that would be necessary if the agent were to be identified with their desire. If *alienation* is characterised in this positive way, there are two possibilities for characterising identification. The first is to also give a positive characterisation of identification as well—perhaps in terms of the desire's relation to the agent's apparent reasons, perhaps in terms of its integration with their other desires. This would leave open the possibility that the alienation–identification distinction is neither exclusive nor exhaustive: nothing in the account would guarantee that someone cannot identify with a desire from which they are also alienated, or that someone can experience a desire from which they are not alienated but with which they are not identified. This is not a very attractive prospect. We arrived at the notion of identification just by contrasting it with alienation. Not only have we seen no evidence of any third alternative relation between an agent and her desire, but it is not clear what this might be, except perhaps a kind of ambivalence between alienation and identification. The simpler and more attractive approach is just to say that, alienation being the positively characterised notion, identification with a desire is just the way one relates to a desire from which one is not

alienated. If that is right, then there is no need for an explanation of why Hampshire's collector identifies with his desire, except to say that nothing causes him to feel alienated from it. This allows us, in a way the positive accounts of identification did not, to make sense of the idea that his desire can be a real source of reasons for him, even if it does not fit neatly into some pre-existing psychological or evaluative structure. His desire, his reason, can be truly particular, and no less significant for that.

5.3 Desirability characterisations and the question 'Why?'

I have argued that the demand for an account of the collector's identification with his desire is misplaced. He is identified with his desire simply because it is his. No more needs to be said on the matter. This addresses the Watson-style argument for a cognitivist view. If we turn our attention back to the argument we found in Anscombe, though, there might seem to be an even stronger case for thinking that the collector's desire must depend on his taking himself to have some worldly reason to buy the bronze. The argument is that whenever someone wants something, we can ask why they want it, or what they want it for. As we noted, if the answer to this question appeals to a further desire, such that the object of the first desire is desired as a means to attaining the object of this second desire, then the question can be raised again with respect to the second desire. This can be iterated indefinitely, and the series of 'Why?' questions, Anscombe suggests, will only be brought to a satisfactory end when the object of the agent's desire is characterised as being in some respect good or desirable. If no such desirability characterisation can be articulated, we cannot understand the person as wanting the thing at all. An object's being in some respect good or desirable, though, looks to be an objective feature of the object that could in principle be a reason for anyone in the right circumstances to want it. So it looks as if Anscombe's argument, if successful, shows that we can only so much as intelligibly ascribe a desire to someone when we can see some kind of rationalisation of their desire. And if there is a rationalisation of the desire, then it seems that whatever rationalises the desire, rather than the desire itself, will be what rationalises any action motivated by that desire.

One response to this line of argument would of course be to reject Anscombe's argument. However, I am willing to accept that the argument is sound. The flaw in the argument as just sketched lies in the assumption (which, notably, Anscombe does not make) that a desirability characterisation must be, so to speak, a desirability *characteristic*: something possessed by the object of desire, independently of its being desired, that makes it desirable. Nothing in Anscombe's argument rules out the possibility of a desirability characterisation's being based in a desire of the the agent's. The nature of this possibility will become clearer when we consider in more detail the role of the discussion of wanting in Anscombe's broader argument, and hence what the significance is of the 'Why?' question as applied to a person's wanting something.

5.3.1 Wanting and desiring

Anscombe's thesis is formulated in terms of *wanting*. She is quite explicit about what she means by this, and, as I will explain, what she does mean by this is not the same as what I mean in talking about *desire*, although there is perhaps some overlap between the two concepts.

Anscombe's book, as its title suggests, is about intention, and she is interested in wanting insofar as it issues in action for which it provides an end. The key role of wanting in her account of intention and intentional action is that it provides the starting-point of practical reasoning, which corresponds to the *end*-point of the 'Why?' series—the final reason-giving answer that brings that series of questions to an end. Anscombe is explicitly not concerned with desire understood as something experienced, for example, because she takes desiring in the experiential sense as being consistent with doing nothing to try to get what one desires (Anscombe, 1963, pp. 67–8). Such notions, she says, 'are not of any interest in a study of action and intention' (Anscombe, 1963, p. 70). Moreover, Anscombe's interest in, and account of, wanting, and her thesis that what is wanted is wanted under the aspect of a desirability characterisation, are intimately connected with other aspects of her account of intentional action, in particular the role of practical knowledge and practical reasoning.

In Anscombe's usage of 'reason for action', the fact that I want something can be a reason for my action in that it can constitute an answer to the special sense of 'Why?' that applies to intentional action and can reveal a part of the teleological order that characterises my acting as I am. My wanting a Jersey cow can be a reason for my going to the Hereford market inasmuch as it provides the point of my going to the Hereford market. However, this feature of Anscombe's account is not in itself especially friendly to the conception of desires as reasons for action that I have been trying to motivate. First, Anscombe's notion of 'reason for action' is relatively thin. The conception of reasons for action I articulated in Chapters 1 and 2 is closer to Anscombe's notion of a premise of practical reasoning, and she is explicit that 'I want' does not in general occur in such premises. The first premise of the practical syllogism, the premise that provides the end of your action, mentions the thing wanted and characterises it as desirable. It does not characterise it merely as wanted. This appears to be connected with the thought, equally troubling for the idea of desires as generating reasons, that an agent is only intelligible as wanting something insofar as they want the thing under the aspect of some good, or under a desirability characterisation. That might lead us to read Anscombe as advancing a version of cognitivism, and to take it that her argument about the intelligibility of wanting shows that desires cannot in themselves provide reasons for action in our sense—because they cannot in themselves characterise objects of wanting as desirable.

However, this would be too quick. As I have said, nothing in Anscombe's account rules out the possibility that in some cases, the way in which the object of the agent's want is characterised as desirable is simply its being characterised as desired. In order to help articulate this possibility, I want to first consider a recent objection to the 'guise of the good' thesis, which might also be taken as an objection to the Anscombean account as I have

presented it.[114] Seeing how the objection fails will help to show how the account I will present works.

5.3.2 Yao on the naturally attractive

There are different versions of the 'guise of the good' thesis, but the general idea is that desiring necessarily involves seeing the object of one's desire as good. If a version of the guise of the good thesis is correct, then it ought to capture the distinction between intelligible and unintelligible action or between intelligible and unintelligible wanting or desire. In a recent paper, Vida Yao argues that the thesis—including a version of it articulated in terms of 'desirability characterisations'—fails to capture this distinction, because there are cases of intelligible desire (which can motivate intelligible action) where the characterisation of the thing wanted is not as desirable—at least, not if this means 'good'—but as, in Yao's terms, *naturally attractive*:

> [A]ll that we need to cite in order to make sense of an agent's desire or action is something that it is intelligible for a human being to be attracted to; and those things that are intelligibly attractive to human beings need not themselves be good or appearances of goodness. Importantly, this is not to say that the agent herself must see the object of her desire as 'naturally attractive to human beings' – why would she care about that? It is, however, to claim that there must be some quality of the thing that she is attracted to, that she represents as a quality of that thing, and that is itself a quality that is plausibly naturally attractive to human beings. (Yao, forthcoming, p. 12)

If Yao's account were correct, this might provide a nice response to the Anscombean argument on behalf of the proponent of desire-based reasons. Anscombe's mistake, on this account, is to think that the answer to the 'Why?' question applied to a person's wanting can only be adequately answered with a characterisation of the object as desirable. This is, perhaps, one kind of answer, but the class of adequate answers is broader—all we need to make a person intelligible as wanting is to see how the object was something naturally attractive to human beings. Yao's primary examples of 'attractiveness characterisations' that are not desirability characterisations are simple experiential properties of the object. She imagines answers to the question 'Why do you want that?' such as 'Because it's so shiny!' or 'Because it's so huge!' and argues that we need not think of these qualities as (apparently) good or good-making in order to make sense of these answers, or for the answers to make sense of the person who gives them as wanting what they want. Importantly, though, such qualities, because they do not characterise the object as desirable, seem not be to the kinds of properties that we would think of as *reasons* for being attracted to, or for pursuing, the object that they characterise. So it is not obvious that answers to the 'Why?' question that merely give attractiveness characterisations can be understood as rationalisations: they explain the desire without giving the agent's reasons for having it.

[114]. The originator of this objection, Vida Yao, quite reasonably expresses uncertainty about whether the argument applies to Anscombe, because it is unclear whether Anscombe is properly characterised as endorsing 'the guise of the good thesis' as it is commonly understood. See (Yao, forthcoming, n. 11).

While Yao's view would in this respect fit nicely with the position I have been motivating, it seems to me mistaken, at least if read as an objection to Anscombe. Crucially, I believe that Yao's 'Why do you want that?' is not the same question as Anscombe's 'Why do you want that?' This difference is due to the ambiguity in 'want' that I noted above. In Anscombe's case, the 'Why?' question asks for the point of what someone is doing; it asks for further specification of their ends and, eventually, a characterisation of those ends that makes them intelligible as ends. In Yao's case, I want to suggest, the 'Why?' question seems to request what is in effect an elaboration of a desirability characterisation that is already implicitly understood, that desirability characterisation being something along the lines of 'I find it attractive', 'I feel a strong desire for it', or 'It appeals to me'.

Yao's attractiveness characterisations do seem to be things of a kind that someone could intelligibly say if asked why they are attracted to a given thing. However, if we are to understand them as doing anything at all, I think we must understand them, as I suggested, as elaborations on or explanations of a desirability characterisation, because the attractiveness characterisations themselves cannot play the role that is required of desirability characterisations: they do not provide adequate stopping points for the series of 'Why?' questions that seeks for the point of what someone is doing.

Suppose, for example, that you tell me you want a 1978 Ford Country Squire, and I ask you why. Let's suppose you answer 'Because it's so huge'. I could quite sensibly ask, 'And what do you want a huge car for?' This would still be Anscombe's 'Why?' question: I would be seeking to understand the point you see in buying a huge car. So an 'attractiveness characterisation', here at least, seems not to provide a stopping-point for Anscombe's 'Why?'-series. Now if, having been asked why you want a huge car, you tell me that it will make you feel safe, or that nothing says 'luxury' like a huge car, or that you always have a lot of stuff to carry around and your current car is too small to carry it, these answers would each show me the point of your buying a huge car. Evidently, though, they all do so by characterising having a huge car as being in some respect desirable. This is something that pointing out the car's hugeness does not do in itself, and this seems to be precisely because hugeness is not in itself desirable, but only insofar as it is useful, say, or pleasant.

5.3.3 Desire as a desirability characterisation

I said that in the kinds of cases I think Yao has in mind, the kinds of answers she envisages—what I have called attractiveness characterisations—could indeed be adequate answers. As we have seen, though, they cannot do the job of Anscombe's desirability characterisations. So what kind of answers are they, and to what kind of question, if not the 'Why?' question that asks for the point of an action?

To address this, I want to make a connection with another kind of question that Anscombe discusses, in different forms, in a few places in *Intention*. It is not the 'Why?' question that seeks the point of the action, or the first premise in the agent's practical reasoning; rather, it interrogates that premise itself. This can be done in different ways

depending on the character of the premise in question. Premises of practical reasoning can for example be subject to ethical challenge, as in Anscombe's 'But why do what befits a Nazi?' (see Anscombe, 1963, p. 72ff.). Sometimes, though, we might simply be looking for a better understanding of the respect in which the agent sees the thing they want as desirable. Anscombe discusses this with respect to the characterisation of something as pleasant:

> Of course 'fun' is a desirability characterisation too, or 'pleasant': 'Such-and-such a kind of thing is pleasant' is one of the possible first premises. 'But cannot pleasure be taken in *anything*? It all seems to depend on how the agent feels about it!' But *can* it be taken in anything? Imagine saying 'I want a pin' and when asked why, saying 'For fun'; or 'Because of the pleasure of it'. One would be asked to give an account making it at least dimly plausible that there was a pleasure here. Hobbes believed, perhaps wrongly, that there could be no such thing as pleasure in mere cruelty, simply in another's suffering; but he was not so wrong as we are likely to think. He was wrong in suggesting that cruelty had to have an end, but it does have to have a point. To depict this pleasure, people evoke notions of power, or perhaps of getting one's own back on the world, or perhaps of sexual excitement. No one needs to surround the pleasures of food and drink with such explanations. (Anscombe, 1963, p. 73)

What this passage illustrates is that, while 'fun' or 'pleasant' are desirability characterisations, an answer to the 'Why?' question that gives one of these is not necessarily immediately intelligible as such: we may want to know more about what the alleged fun or pleasure of the thing is, about how it is fun or what kind of pleasure it involves. A notion of the agent's point of view is again salient here. We want to understand the point that the agent sees in their action, and this means gaining some appreciation of what is good about it from their point of view. When the good of the action or thing pursued is characterised by some relatively objective property that it possesses, such as its being good for human health or being such as to suit the agent's needs, we can, so to speak, occupy the agent's point of view just in looking at the action or object itself. Sometimes 'pleasant' desirability characterisations work like this, as in Anscombe's examples of food and drink. We understand that food and drink just are pleasant. Pleasures, though, can also be idiosyncratic. Some people enjoy things that most of us do not, and sometimes what it is that a person enjoys about the thing they enjoy is not immediately easy for us to grasp. In such cases we may want a richer characterisation of the pleasure in question. Sometimes the best characterisation we can give of a pleasure will be little more than analogical or gestural, and gaining a real understanding of the pleasure in question may be a genuine achievement. Some pleasures, perhaps, can only be fully understood by those who have experienced them. Seeing things from the agent's point of view here seems to require not just attending to the right facts or supposing that things were as they took them to be, but 'getting inside their head' in a deeper way, a way that might require a good deal of imagination.

Now, Yao's attractiveness characterisations are not exactly characterisations of a pleasure, but I do think that we can construe them as playing a similar role in explaining action as do further characterisations of pleasure, and the question to which they are addressed is somewhat akin to the question 'What's the pleasure of it?' Specifically, I think

we can understand them as being addressed to the question 'What attracts you to it?' This is a question that can be expressed in the same words as Anscombe's 'Why do you want that?', but it is somewhat different, as I will explain.

Let's return to our example. I ask why you want a 1978 Ford Country Squire, and you reply 'It just appeals to me'. This, I am claiming, can be a desirability characterisation. The respect in which you see having that car as desirable is that you feel drawn to it; it speaks to you; you find it appealing. In other words, you desire it. Now, I might find this desire of yours puzzling, particularly if I myself find this boxy car, with its synthetic wood grain exterior, wholly unappealing. So I might ask, 'Why do you want it?', meaning, roughly, *What about it appeals to you?* Here Yao's 'Because it's so huge!' might be an informative answer. If you answer in this way, I can try somewhat better to imagine how it is that you find this ugly car appealing. Something about its size appeals to you. Insofar as the 'attractiveness characterisation' makes your desire more intelligible to me, though, this does not seem to be because I appreciate the hugeness of the car as something that would be a reason for anyone in your circumstances to want to buy the car. It is simply that it is, as Yao suggests, the kind of thing that might conceivably attract someone. It is only because it has this effect on you, though, that you have the reason that you have to buy the car. The reason depends upon your actually being attracted to the car.

Note that if, in this case, when you explain that you are attracted to the car because it is so huge, I ask you what the point is of having such a huge car, I seem to have misunderstood you. In this, the case differs from that in which 'Because it is so huge' was offered as an answer to the 'Why?' question that seeks for the point in what the agent wants. The point in having what you want, here, has already been established: you have a desire for the car; it appeals to you. Its hugeness is simply something you appeal to in trying to characterise the appeal that it has for you.

Note also that it is not clear that you must actually be able to say anything as informative as a Yao-style attractiveness characterisation in order for your desire to give you a reason to pursue its object. In the case of Hampshire's collector, for instance, I think we can perfectly well imagine him having little to say about why he desires the inferior bronze—he might well say something like 'I don't know why I want it, it just appeals to me', or 'There's just something about it'. Such answers do not make his wanting the bronze unintelligible. They simply reveal something about what kind of desire is at play here: it is the kind that its subject simply feels, and for which they have, and need, no further account.

5.4 How can desire provide a desirability characterisation?

Anscombe's argument for the claim that whatever is desired under the aspect of some desirability characterisation seemed to pose a threat to the view that a desire can, in a case like that of Hampshire's collector, provide a stopping-point for rationalisation of an action. The thesis about desirability characterisations suggests that when we explain an action in terms of a desire, there is always more to be said. Moreover, it seems to suggest that what is

left to say is something bringing the agent's desire under the schema of universal reasons, since what is required is that the desired object be characterised as instantiating some form of objective good. I have argued that this apparent conflict between Anscombe and Hampshire can be resolved if we acknowledge that one way of characterising something as desirable *for you* is to explain that you desire it. This suggestion might seem puzzling. I seem to be saying that a person's wanting something can be explained by their wanting it, which sounds nonsensical. Although I have already pointed out that there seems to be an ambiguity in 'want' at play here, I should probably say more about how to make sense of the claim that desire can provide a desirability characterisation.

First, recall the observation that 'want' and 'desire' are ambiguous. I am by no means the first to make this point.[115] In one sense, 'desire' is roughly synonymous with 'motivation', so that any intentional action manifests a desire, which provides the aim with which the agent acts. In the second sense, a desire is something distinctively passive,[116] perhaps characteristically 'affective'[117] or consciously felt. This is the kind of state we are talking about when we talk about feeling *attracted* to someone or something, or something's *appealing* to you (or your *finding* it appealing), your feeling *drawn* to some object or activity, and so on. Desire in this narrower sense is richer than the more abstract or purely logical notion that can be applied to every intentional action: it is a state of mind with what Wollheim calls *psychological reality*; it is a thing of substance, which arises at a time and develops over time, which has a kind of life-cycle or natural history, and which shapes the subject's conscious experience. As I explained above, Anscombe's argument about the 'Why?' question and desirability characterisations concerns the former, broader, more abstract notion of desire. What I am suggesting is that one's desiring in the narrower sense can explain one's wanting in the broader sense. (I will henceforth use 'want' for the broader concept and 'desire' for the narrower.)

There are two possible ways to flesh out this explanatory relationship. The first is to see the desiring as something that in some way stands behind the wanting, so that one's desiring what one desires is the basis on which one is motivated to pursue it. Perhaps the desire causes the wanting; perhaps it is the reason for the wanting. On the second approach, desire is itself a motivational state, so that desiring essentially involves wanting. If this is right, then the desirability characterisation 'It appeals to me' or 'I desire it' does not explain one's wanting by appeal to an independent mental state; rather it recharacterises one's wanting as, so to speak, a desirous wanting. It says: this is the kind of wanting that simply comes over one, not the kind that is based on some further reason or motive.

Compare love. Love, many authors hold, essentially involves certain characteristic motivations, perhaps most notably the motivation to act in the beloved's interest.[118] Being so motivated, on this view, is a part of what it is to love. Whether or not this view is correct, it would be a mistake to object to it on the basis that it rules out the possibility of informatively

[115]. See for example (Davis, 1986; Nagel, 1978; Schapiro, 2014; Schueler, 1995).
[116]. (Schapiro, 2014).
[117]. (Chang, 2011).
[118]. See for example (Frankfurt, 2004; Kolodny, 2003; Taylor, 1975).

answering 'Why do you want to help him?' with 'I love him'. It does not rule that possibility out; it simply gives a specific account of what the envisaged answer does and does not say.

Indeed, in the next and final chapter, we will turn our attention from desire of the sort that has been our focus since Chapter 3 to look instead at love. Hampshire describes the collector as '"falling in love", as we say' with the bronze, and while love and desire are two different things,[119] there are significant parallels between falling in love and forming a desire. Moreover, I think that by reflecting on love, we can get a clearer picture of how we can make our actions intelligible to others when we do not act on a universal principle. Looking at love might thus help, albeit indirectly, to make clearer the kind of explanation of action that a desire can provide. It will also, I think, help to explain what I have called desire's psychological reality. And it will help to show that the kind of personal, idiosyncratic reasons illustrated by the example of Hampshire's collector are more diverse and include reasons of more significance than mere whimsical attractions or passing fancies.

[119]. (Holloway, 1966).

Chapter 6

Love is Weird

Seems love ain't the way that it oughta be
It tends to depend too much on anatomy
But I suppose that's all well and fine
I'll be yours if you'll be mine

– Daniel Johnston, 'Love is Weird'

6.1 Is love a rational attitude?

Love and desire are both, as we might say, modes of caring. They each, in different ways, assign a personal significance to things that would otherwise lack it. By 'personal significance' I mean a significance for the possessor of the attitude: significance, specifically, to the lover or to the desirer. Love, like desire, seems to rationalise actions. We do things for those we love that we would not do if we did not love them. As in the case of desire, we can ask whether this rationalising role is fundamental, and as in the case of desire, we can connect this question to questions about the relationship between love and reasons. Do we love for (apparent) worldly reasons? Is love rationalised by considerations that would be reasons for anyone in the lover's situation to love as the lover does? Is it the reasons for love, rather than love itself, that make sense of loving actions from the lover's point of view? If so, what are those reasons?

In the case of desire, the equivalent of the last question—what the reasons for desiring would be, if we do indeed desire for reasons—invites a fairly straightforward answer: the reasons would be the same as the reasons for getting or doing whatever the desire is a desire to get or do. This straightforward answer is possible because we tend to think of a desire as a fairly simple motivational state, individuated by its object, in such a way that the psychological role of any desire is just to motivate its subject to try to attain its object. The connections between love and motivation are more complex, and it also seems to have very

important and equally complex connections to emotion. Love relates a person to particular person or object, but whereas the role of desire for an object is (at least in the first instance) quite obviously to motivate the agent to get its object, the actions love motivates involve the beloved in many different ways and sometimes barely involve the beloved at all.

This has led to interesting discussion, among those who assume that love is based on reasons, about what the reasons for love are. In this chapter, I will explore some of the arguments for one influential account of the reasons for love. I will argue that, while that account is in important ways close to the truth, it presents an inaccurate picture of how we make sense of love from the lover's point of view. In seeing how it does so, we can, I think, begin to see our way to a different way of thinking about the intelligibility of attitudes like love and desire, both in terms of the sense they make to the subject and in terms of how they are made intelligible to others.

6.2 The quality theory

Insofar as there are reasons for wanting, they seem in general to be facts that show the potential object of wanting to be in some respect worth pursuing, doing, having, engaging in or bringing about. They concern the goodness or desirability of the object itself. Since love, like wanting, is broadly speaking a positive or favourable attitude, it would be natural enough to suppose that insofar as there were reasons for love, they would be considerations of the same general kind: considerations of the goodness or value of the potential beloved. Think of one traditional kind of wedding speech. The groom, apparently attempting to explain why his new wife is so special to him, lists as many of her wonderful and charming qualities as he can: her intelligence, her sense of humour, her kindness, her beautiful smile, and so on. On the most straightforward version of what has come to be called the *quality theory*, the reasons for love are just these kinds of facts about the beloved. However, if we try to think this suggestion through, it turns out not to give a very plausible picture of how we actually make sense of love.

The general shape of the problem starts to become apparent when we consider the outsize role that the people and things we love play in our lives, in our practical deliberations and our emotional responses to events. Loving someone involves caring a great deal about what happens to them, and therefore giving their interests a special relevance or priority in your thinking about what to do. As it is often put, love involves a form of partiality to the beloved. The people you love are, in general, more important to you than those you do not love. This partiality, this special importance to the lover of the beloved, is something that makes sense to the lover. It does not strike the lover as strange or irrational or unjustified.[120] However, this partiality seems capable of comfortably coexisting with an awareness of the beloved's shortcomings and of the fact that for any given good quality that they possess there

[120]. At least, as long as the lover has not engaged in certain kinds of philosophical reasoning. See (Williams, 1981b) for a classic discussion, and criticism, of some of the ways in which ethical theories can make this kind of partiality seem more dubious than we ordinarily take it to be.

might be other people who possess the same virtue to a greater degree. As Harry Frankfurt puts the point, we

> commonly think that it is appropriate, and perhaps even obligatory, to favor certain people over others who may be just as worthy but with whom our relationships are more distant. Similarly, we often consider ourselves entitled to prefer investing our resources in projects to which we happen to be especially devoted, instead of others that we may readily acknowledge to have somewhat greater inherent merit. (Frankfurt, 2004, p. 35)

Of course, we typically focus on the good in the people we love and do not focus on the bad, and we often see good in them where others might not. This is a common, perhaps essential, aspect of loving. The fact remains, though, that the priority in personal importance accorded to those we love does not in any straightforward way map onto an objective scale of evaluation. We do not only love those we think best or most worthy; or, if we do think of those we love as the best or most worthy, this seems to us to be a value or worth that is bestowed on them by our love, not an objective quality to which our love is a response.[121] The kind of traditional wedding speech I described above, if it is really meant to be an explanation of the groom's love, is incredibly facile. The qualities he lists are ones any number of other people could have, but he is appealing to them to account for what is supposedly a deep and unique attachment to this particular person. The 'reasons' do not seem to fit the nature of the attitude.

This line of thought can be sharpened into a number of distinct but connected objections to the quality theory. We can begin with the following four:

- *Universality*: If my reasons for loving you are qualities you have, should anyone else who is aware of those qualities love you too and in the same way?
- *Promiscuity*: Should I love, in the same way, anyone else who has the same qualities?
- *Trading up*: If someone else has the same qualities to a greater degree, should I love them instead, or more?
- *Inconstancy*: If you lose the relevant qualities, should I stop loving you?[122]

It seems clear enough that the answer to each question should be negative.

To an extent, most of these objections can be accommodated if we simply accept that love is not *maximising*, that is that we are not required to love all and only the best people, or to prioritise our love according to some ranking of people on a scale of personal quality. The reasons for love, we can say, are permissive or 'noninsistent', so that a person's good qualities 'recommend' loving them, or make them 'eligible' for love, but do not require anyone to love them. This gets us around Universality because even if my love for you can be justified only if everyone has some reason to love you, it does not follow from the fact that everyone has reason to love you that everyone ought to love you. Similarly, Promiscuity is avoided because the valuable qualities that justify my love for you do not require me to love you and so do not require me to love anyone else either. Finally, for the same reason, we also have an answer to

[121]. (Frankfurt, 2004).
[122]. Adapted from (Setiya, 2014).

Trading Up: since the reasons for loving do not require me to love, someone's being 'better' or 'more lovable' than the one I love does not require me to love them as well, or instead.

Saying that love is not maximising does not in itself answer the problem of Inconstancy, but it makes a fairly straightforward response possible. Everyone, or just about everyone, presumably has some good qualities. If love is not maximising, perhaps it does not take very much for love to be justified; it is enough that the actual or potential beloved have some good qualities. If so then the kind of change necessary for one to lose one's justification for loving someone might have to be really quite extreme—they might have to lose essentially all of their good qualities, becoming some kind of monster. In such a case, perhaps one really ought to stop loving them.

While we can in this way respond to the letter of the above objections, the response still is not entirely satisfactory. If the reasons for loving a particular person are noninsistent, so that you can love someone rationally and justifiably despite having equally good or even better reasons to love someone else, then your reasons do not determine your love for the one you love. With respect to the reasons that apply to you, your loving this person and not some other is completely arbitrary. From the lover's perspective, though, its being *this* person who you love, who is so important to you, is not arbitrary in this way. They are not simply chosen at random from among the available options, and it would not be reasonable or even intelligible for you to simply decide to replace them with someone else.

6.3 The relationship theory and the particularity of love

The deep problem for the quality theory stems from the particularity of love. This can be illustrated by another objection, which Niko Kolodny calls the problem of *Nonsubstitutability*:

> If Jane's qualities are my reasons for loving her, then they are equally reasons for my loving anyone else with the same qualities. Insofar as my love for Jane is responsive to its reasons, therefore, it ought to accept anyone with the same qualities as a substitute. But an attitude that would accept just as well any *Doppelgänger* … that happened along would scarcely count as love. (Kolodny, 2003, pp. 140–1)

Note that seeing the reasons for love as permissive or noninsistent does not resolve this puzzle. The thing the quality theory cannot explain is not that I am not required to love Jane's *Doppelgänger*; it is that I have reason to love Jane *instead of* the *Doppelgänger*, and that if I was willing to accept the *Doppelgänger* as a substitute this would show that I did not really love Jane at all. The quality theory cannot in principle explain the character of love as an attachment to a particular person because of the way in which it seeks to explain love in terms of something essentially general, namely qualities which could in principle (and often in practice) be instantiated by someone other than the beloved.

Kolodny also raises further objections which provide further support for the idea that the particularity of love cannot be explained in terms of qualities of the loved one. There is the problem of *familial love*: we typically love people to whom we are closely related, and to

explain our love for these people we would most naturally appeal to our familial relation to them rather than to their personal qualities. There is the problem of *modes*: there are different ways of loving, and the way in which it is appropriate for you to love a given person depends on something other than their personal qualities. One can love someone as a friend, for example, or as a romantic partner, or as a child or sibling or parent, and it is often the case that it is appropriate for different people to love the same person in different ways. If Heather's mother ought to love her in a different way from her best friend, this difference must be explained by something other than Heather's personal qualities, since these are accessible to both the mother and the friend.

Finally, there is the problem of *amnesia*. Plausibly, losing certain memories could cause you to lose your love for someone. If love is a response to reasons, it is natural to think that this is because in losing the relevant memories, you lose access to the reasons for you to love the person. However, we can imagine a case wherein loss of memory leads to loss of love, even though the amnesiac lover remembers everything about their erstwhile beloved's personal qualities. Kolodny illustrates this with the story of the 'amnesiac biographer':

> [The biographer] spent his early fifties writing the biography of a contemporary, a political activist whose accomplishments were already noteworthy by that age. His biography drew on the reminiscences of her closest friends and amounted to a strikingly intimate portrait of her life and character. As a result, he found her in many ways admirable and attractive, but they had never met, and the thought of a relationship with her never entered his mind. ... In their late fifties, they met, fell in love, and married. ... A decade later he suffers a special kind of memory loss. He can recall everything that happened to him up until a few years before their relationship started, but nothing after. (Kolodny, 2003, p. 141)

Kolodny claims that we 'would not expect him to love her, and indeed it is hard to see how he could', because '[t]o him, she is no longer the woman he fell in love with' (Kolodny, 2003, p. 141). Since he is still well aware of all her 'attractive and admirable' qualities, this suggests that it is the relationship between the two, and not her qualities, that formed the basis of his love for her.

Kolodny proposes an alternative account of the reasons for love that apparently resolves all of these worries. On his *relationship theory*, the reason for which one loves, when one loves someone rationally, is one's relationship to the person one loves—or, to be more precise, it is the fact that one has a relationship with the relevant person, where that relationship is of a finally valuable type.[123] Such relationships are, Kolodny says, constituted by patterns of interaction marked by mutual noninstrumental concern and emotional vulnerability. Roughly, the idea is that my love for my wife makes sense because of all the things we have done together and the way in which our doing those things connected with our feelings about each other and concern for one another. A relationship like this is a valuable thing and its value makes sense of my valuing both my wife and also the relationship itself. This valuing —valuing the relationship and the one with whom one has that relationship—is what, on Kolodny's accont, love consists in. 'Valuing' is further analysed as consisting in certain beliefs,

[123]. This is made explicit on (Kolodny, 2003, p. 151).

emotional dispositions and standing intentions. In detail, Kolodny defines A's loving B as consisting in A's:

> i. believing that A has an instance, r, of a finally valuable type of relationship, R, to person B (in a first-personal way—that is, where A identifies himself as A);
>
> ii. being emotionally vulnerable to B (in ways that are appropriate to R), and believing that r is a noninstrumental reason for being so;
>
> iii. being emotionally vulnerable to r (in ways that are appropriate to R), and believing that r is a noninstrumental reason for being so;
>
> iv. believing that r is a noninstrumental reason for A to act in B's interest (in ways that are appropriate to R), and having, on that basis, a standing intention to do so;
>
> v. believing that r is a noninstrumental reason for A to act in r's interest (in ways that are appropriate to R), and having, on that basis, a standing intention to do so; and
>
> vi. believing that any instance, r*, of type R provides (a) anyone who has r* to some B* with similar reasons for emotion and action toward B* and r*, and (b) anyone who is not a participant in r* with different reasons for action (and emotion?) regarding r*. (Kolodny, 2003, p. 151)

Recall the objections to the quality theory. Universality, Promiscuity and Trading up simply do not arise. Inconstancy is not a worry either: the beloved can go through all kinds of changes; as long as they have a valuable relationship with the lover, love ought to remain constant. Familial love is explained by the fact that familial relationships are of a relevant finally valuable kind. Different kinds of relationship (parent–child, friendship, romantic relationships ...) constitute reasons for different kinds of love, addressing the problem of modes. The amnesiac ceases to love his wife because he forgets all about their relationship, which was his reason for loving her. Finally, love does not accept substitutes—it is particular —because no matter how intrinsically similar a *Doppelgänger* is to one's beloved, it is one's beloved with whom one has a finally valuable relationship, not the *Doppelgänger*.

The relationship theory answers the challenges to the quality theory by thickening the basic cognitivist conception of reasons in certain ways. What the most serious challenges to the quality theory were getting at was the tension between, first, the character of love as an attachment to a particular person and, second, the character of qualities as something essentially general. The relationship view resolves this tension by taking the reasons for love to be essentially *relational*: love is justified not just by the beloved's being a certain way, but by the lover's having a certain connection to the beloved. Moreover, what anchors us to those we love as particular individuals is our *history* with them—a history which is itself a series of particular events, involving particular people. Making sense of love, and of the actions that love motivates, from the lover's point of view thus involves more than just considering what the lover takes to be features of the options open to her in a way that might be suggested by a crude cognitivist view. What it makes sense for one to do is not just a

matter of what is objectively good or worthwhile: it is also partly determined by how one's history relates one to good and worthwhile things.

A further complexity introduced by the relationship theory is a certain kind of indirectness in the rationalisation of loving actions. In the case of desire, there is a straightforward relationship between the reasons that according to the cognitivist rationalise desire and the actions that desire motivates: the reasons concern the goodness or desirability of the object that desire motivates the agent to get. In the case of love, on the relationship view, many of the lover's actions and emotions concerning the beloved are in fact rationalised by the value of something else, namely the relationship. As Kolodny explains his theory, love consists in valuing both one's beloved and one's relationship with one's beloved, but the valuing of both is rationally grounded just in the value of the relationship. One's valuing attitude thus has two 'foci' but one 'ground', and while one values the relationship *finally*, in that one sees it as the source of one's reasons for valuing it, one values the beloved non-finally but also non-instrumentally, in that while one does not value the beloved merely as a means to some further end, one nonetheless sees something other than the beloved (namely the relationship) as the source of one's reasons for valuing the beloved. Hence when the lover acts to, for instance, benefit or protect their beloved, the rationalisation of this action goes deeper than simply citing considerations that would be reasons for anyone in the lover's circumstances to act in that way (unless we operate with a very rich conception of 'circumstances') and there is a somewhat indirect connection between a person's loving actions and the value that grounds the reasons that fundamentally rationalise those actions. The value of your relationship with your beloved is fundamental to the intelligibility of your concern for them, but it will nonetheless be the case that many of the loving actions you perform, which express your concern for your beloved, are not performed in order to 'promote' the value of that relationship.

Finally, we should note that some relationships, such and friendships and romantic relationships, are *attitude-dependent*. These kinds of relationships are partly constituted by what Kolodny calls 'patterns of concern'; that is, a given historical relationship between two people is a friendship only if it involves certain kinds of emotional responses that reflect each party's concern for the other. The historical, relational reason for loving in such cases thus involves a distinctively 'subjective' element: one's own past emotional responses to this person.

Despite all this, Kolodny's relationship theory is clearly and resolutely cognitivist. Although the relationship that justifies love needs to be understood in relational, historical, and (in part, in some cases) affective terms, it makes sense of love just because it is a valuable kind of relationship. A relationship of such a kind is a reason for anyone to respond in the same way—by valuing the relationship and the person to whom they are so related. Moreover, as Kolodny's conditions (i)–(vi) make very clear, loving consists in large part in taking oneself to have various kinds of universalisable reasons. And, crucially, love is rationalised by the *belief* in, or knowledge of, the fact that one has a relationship of a finally valuable kind with the beloved.

6.4 Relationships and the lover's point of view

To reiterate: while Kolodny's slogan is that the reason for love is a valuable relationship, his view is not that the reason for love is a certain historical particular—the process or event or series of events we call a relationship—but a certain fact. 'The reason for love is a valuable relationship' is, in his account, shorthand for the claim that the reason for love is the fact that one has a relationship of a valuable kind with the other person. Specifically, where person A appropriately loves person B, A's reason for loving B is the fact that A has relationship r to B, where r is an instance of a finally valuable relationship-type R. A's love is rationally based on her belief that she stands in r to B, and also includes, on Kolodny's account, certain beliefs about what kinds of responses the fact that they are so related is a reason for.

The focus on the historical relationship between lover and beloved is a major strength of Kolodny's account. It allows him to explain the way in which love is highly selective without claiming, implausibly, that it is typically based on an assessment of the beloved as being in some way better than every other potential object of love. The view seems to capture a feature of love that might initially seem quite puzzling, namely the way in which one's love for a particular person is manifestly highly contingent, in that one could easily have ended up loving a different person instead, without being, from the lover's perspective, arbitrary. While I think he is right to focus on the historicity and relationality of love, though, I also think that, in maintaining a broadly cognitivist picture of rationalisation, Kolodny misconstrues the psychological role of the historical and relational factors. The psychological significance of love's history goes deeper than the rationalising role of the lover's recognition of a single fact. This can, I think, be brought out by considering in more detail one of Kolodny's arguments for the relationship view.

Recall the case of the amnesiac biographer:

> [The biographer] spent his early fifties writing the biography of a contemporary, a political activist whose accomplishments were already noteworthy by that age. His biography drew on the reminiscences of her closest friends and amounted to a strikingly intimate portrait of her life and character. As a result, he found her in many ways admirable and attractive, but they had never met, and the thought of a relationship with her never entered his mind. … In their late fifties, they met, fell in love, and married. … A decade later he suffers a special kind of memory loss. He can recall everything that happened to him up until a few years before their relationship started, but nothing after. (Kolodny, 2003, p. 141)

Kolodny says that in this case, we would not expect the biographer to love his wife, and 'it is hard to see how he could'. Since he has only forgotten their relationship, and not her valuable qualities, this tells against the quality theory in favour of the relationship theory. However, if we consider some variations on the example, we can also raise some explanatory challenges for the cognitivist relationship theory.

6.4.1 First variation

Imagine that, after his memory loss, the biographer's partner tries to remind him of their relationship. She shows him various kinds of of documentary evidence—videos and photographs of them together, love notes, instant message threads, and so on. She tells him stories of their time together, good and bad. She hopes to jog his memory, to get him to recall their relationship. She fails—his memories are lost for good—but he accepts the evidence and trusts her testimony, and thereby comes to know that they had a deep, loving relationship, and comes to know many details about things they did together and their feelings about each other. Kolodny's relationship view suggests that, given this knowledge of their relationship, he ought to love her: it would be inappropriate, indeed irrational, for him not to. By the same token, it would make sense, from his point of view, to love her.

This seems to me the wrong verdict. It is not so much that I think he ought *not* to come to love her as a result of being given all this evidence. Rather, it is that whether we can understand him as genuinely coming to love her depends on much more than just his regaining cognitive access to what is, according to the relationship view, *the* reason for him to love her. If love is rationalised as the cognitivist says, by a fact or a belief in the obtaining of a fact, then all that should be required for love to make sense from the biographer's perspective is that he have the fact in question in view, that it be part of his picture of how things stand. But this alone does not seem to make sense of love.

There are ways in which we might find intelligible the biographer's coming to love his wife again after his memory loss. One would be if we could understand him as falling in love with her all over again. Whether we (and, more importantly, he) could understand that would depend on the interactions between the two of them here and now, post-amnesia: how they get on, what feelings are sparked, and how those feelings develop. His learning about their past relationship might play a role in triggering and strengthening those feelings, and hence his learning about the past relationship might play a significant role in getting him to come to love her again. This would certainly be an emotionally significant thing to learn. However, the emotions that it might immediately trigger would not, I think, constitute love. And insofar as he did come to love her again in this way, this is not a story of his regaining his old love when he regains knowledge of that love's rational basis. It is a story of his being charmed by her all over again, much as he was when they first met. This is a plausible enough thing to imagine—after all, it happened once before. Any love that developed in this way would, though, at least for him, have the character of a *new* love, and not the mature love which, on the relationship view, a mature relationship would make appropriate.

6.4.2 Second variation

Moreover, note that things need not go this way. Suppose that when our biographer comes round, his wife seems to him a stranger. He finds her interactions with him overfamiliar and he feels suspicious. Even when he comes to learn about their relationship, he feels no real warmth towards her. He does not understand 'from the inside' how he came to love this

person, in part because he cannot remember 'from the inside' the experiences in and through which he originally did so. He may feel remorseful about his lack of feeling, but it is not unintelligible to him. Because he cannot recall first-personally or episodically any of the events that she recounts, they do not have for him the emotional force that they would have had before he lost his memory. He believes the stories he is told, but they leave him cold and alienated.

On the relationship view, this constitutes a failure of rationality. The biographer has regained access to a conclusive reason for loving this person, and he is failing to respond to it. Again, this seems the wrong verdict: his loss of feeling is certainly unfortunate, and it is certainly a case of things in his mind not going as they ought to. To call it irrational, though, is to misdescribe the kind of 'going wrong' that it constitutes. His loss of feeling is no more a failing of rationality than is his loss of memory. Indeed, it is a sad consequence of the latter loss, or of whatever caused that loss.

6.4.3 An unattractive response

One line of response a defender of Kolodny's relationship view might try to make here is to propose a restriction on the way in which the relevant fact must be known in order to play the role of rationalising love. This does not seem to me a very attractive route to go down. If we have no reason, other than the objection I have raised, to think that there is such a restriction, the suggestion seems objectionably arbitrary. To illustrate this point I will consider one possible such restriction.

One possibility is that a reason for love can only play its distinctive rationalising role if it is known about in a way that is immune to error through misidentification (IEM).[124] After all, Kolodny says, in his condition (i), that A must believe 'that A has [a finally valuable relationship] to B (*in a first-personal way—that is, where A identifies himself as A*)' (Kolodny, 2003, p. 151 emphasis added). This kind of first-personal belief is commonly thought to be IEM. However, the kinds of information-source through which the biographer regains his knowledge post-amnesia are not essentially first-personal, and hence are not essentially IEM.

Let me explain. Before the onset of his amnesia, the biographer knows of his relationship with his partner through experience and memory. His judgment that he is in the relationship being based on this kind of information-source, the following kind of error is not possible: he knows that someone is in the relationship, but mistakenly thinks it is someone else. However, that kind of error is in principle possible when his judgment is based on the evidence his partner gives him after he loses his memory. In the latter case, he needs to take the 'extra step' of identifying himself as the person in the relationship. In this respect, he does not, in our modified amnesia case, know about the relationship in an essentially first-personal way.

The requirement that the lover should know about his relationship with the beloved in an essentially first-personal way—basically, that the lover must know the relationship from

[124]. This terminology originates in (Shoemaker, 1968).

his own experience and memory—seems to give the right verdict regarding our variations of the amnesia example. However, it is not clear that the cognitivist has the resources to explain why there should be such a restriction on the rationalisation of love.

Consider one possible motivation for the importance of first-personal, IEM belief in cognitivist reasons. It is widely held that first-personal thoughts, those naturally expressed with 'I', have a special significance for agency.[125] Such thoughts seem to be IEM. Perhaps, then, the biographer needs to recognise first-personally that he himself is or was in a valuable relationship with the person who is his partner if that fact is to have the right kind of emotional and practical significance for him. It would not be enough if he merely came to know that some man was in such a relationship, where, as it happens, that man is in fact him. This is why Kolodny specifies that A must know in a first-personal way that A is in the relationship.

The problem is that the restriction that this perfectly sensible line of thought motivates is not strong enough to explain what is going on in our variations on the amnesia example. Those variations as described meet all of the relevant conditions: the biographer does not only know that some man had a relationship with this woman, where as it happens that man was him. He knows that that man is he himself. He thus knows in a first-personal way that the relevant fact, which according to the relationship theory is the reason for him to love this woman, obtains. He meets the condition because he (first-personally) identified himself as the person who was in the relationship. What is needed, then, is the much stronger restriction that his way of knowing that he was in the relationship is IEM. This is the condition that fails to be met in our variations on the amnesia case, because testimony and documentary evidence about oneself are not IEM in this way. However, this stronger restriction seems unmotivated. Compare affective responses other than love, such as anger. To be angry at someone for wronging you—at least, to experience a particular form of anger—it is plausible that you need to know (or think) that they wronged you yourself, that is to say that the belief on which the anger is based must be first-personal. But do you need to know about the event of wronging in a way that is IEM? It seems not. If you learn of the wrong later on, by testimony or documentary evidence, the indirectness of your knowledge of the event will not in any way prevent your knowledge of the wrong from rationalising your anger. It is unclear why love, if it is a rational attitude, should be so different.

6.4.4 A third variation

With the first two variations in view, we can recognise a third possibility, happier than either of the first two. Rather than losing his love and either developing a new love or not, it seems conceivable that the biographer's *old* love for his partner might survive whatever trauma caused his memory loss. That is, the biographer might feel all (or at least very many) of the characteristic emotions and motivations of a mature love for this person, even though he

[125]. Widely but not universally held. See (Perry, 1979) for the main source of the contemporary debate. (Cappelen & Dever, 2013) reject the special significance of 'I'-thoughts. (Babb, 2016) provides one response.

cannot remember the details of their relationship together. On the relationship view, such feelings would be irrational; they would not make sense from the biographer's point of view. Again, though, whether that is the correct assessment seems to be a question of detail.

Suppose that, although he cannot remember anything factually about their time together, the biographer, when reunited with his wife, feels a familiarity and warmth towards her, and even feels that he is deeply in love with this woman, whoever she may be. He lacks access to an explanation of these feelings—although he might, on the basis of the feelings themselves, have an idea of what the explanation must be—but they might nonetheless 'make sense' from his point of view, if an attitude's 'making sense' to its subject is essentially opposed to his feeling alienated from it or finding it unintelligible. His feelings, we might say, feel right to him. Not only is this case conceivable, but something like it might have really happened to none other than Derek Parfit. Parfit, apparently experiencing a kind of nervous breakdown, temporarily lost his memory of much of his past, including his marriage to his wife, Janet Radcliffe Richards. Nonetheless, when asked if he knew who Richards was, he reportedly replied: 'Yes. She's the love of my life.'[126]

This third case on its own would not constitute such a strong challenge to Kolodny's view. On the one hand, Kolodny could insist that the biographer's feelings do not make sense from his point of view until he learns about the relationship. There is perhaps some truth here. The biographer does not, in this case, have access to a story of the kind that would explain his feelings. However, if the claim is that he would necessarily feel alienated from his feelings, this is exactly what I am denying. Recall the discussion of alienation in the previous chapter. It is entirely possible that something similar could be at play here. The biographer feels something that just makes sense, in itself. He will only feel alienated from it if something leads him to reject or disavow his feelings. I suppose we can imagine a case in which the biographer, disturbed by his lack of explanation for his feelings, does distance himself from them, and does therefore feel alienated from his love for his wife. But we can just as well imagine that he simply trusts himself and accepts those feelings as being truly his own.

There is a more promising response for the defender of the cognitivist relationship theory. This is to insist that, insofar as the biographer's feelings are intelligible to him, he just knows—perhaps even knows on the basis of his feelings—that he has a valuable relationship with this person. Parfit's statement that Richards is the love of his life expresses, on this account, not just the strength of his present feeling, but an incipient, if perhaps not fully explicit, recognition that this is someone with whom he has a certain sort of history. Endorsing this response would involve taking on certain epistemological assumptions, but it is those assumptions are not obviously implausible. So there may be a way for Kolodny's relationship view to account for this third variation on the amnesia case. In the light of the

[126]. (MacFarquhar, 2011). It is unclear whether the case as it actually happened exactly fits the structure of my revision of Kolodny's vignette, but it is at least conceivable that it might have done.

case already made against the rationalistic character of that view, though, we can reasonably doubt whether we must describe the case in this way.

6.5 How does a relationship figure in the lover's psychology?

If what I have argued is correct and the rationalistic relationship view gives incorrect verdicts about when love is intelligible, what should we conclude about love? One response would be to look for an alternative account of 'the reasons for love', of what kind of fact, what kind of feature of the subject's perspective on how things are, makes sense of love. I want to recommend a different response. I think the proper diagnosis of the relationship view's failure is not merely that it picks the wrong kind of fact as love's justifying reason. The mistake runs deeper: it concerns the character of love as an attitude and the kind of explanation to which it is subject.

Because of its cognitivist character, Kolodny's relationship theory fails to do justice to two important and interrelated features of love—features that also characterise the kind of desire that we discussed in Chapters 3–5. These features have significant implications for how we should think about the relation between making one's motivations or actions intelligible to oneself and others and there being a universal principle under which one's feelings or actions can be subsumed.

The first feature is love's subjectivity: love is grounded, at least in part, in the lover. It is not simply a response rationally determined by universal principles or values external to the individual lover, and it is not the case that, insofar as I love appropriately, anyone in my circumstances ought to love as I do. As Kolodny's view stresses, in loving one is related to another external to oneself. This relation is itself an objective fact, and the relationship may be, as Kolodny claims, objectively valuable. Love might nonetheless be subjective in that this relation itself has its ground, in part, in contingent features of the lover, as well as in features of the beloved and their historical interactions. This holds, moreover, from the lover's perspective—or at least it can, from the perspective of a reflective lover. Your love need not be undermined by your recognition that it largely arises from a brute, non-rational susceptibility to be affected in a certain way.

Secondly, love is psychologically real. Love is not abstract in the manner of the notion of 'wanting' discussed in Chapter 5. A love is something with substance, which emerges and develops over time. It has a natural history.[127] This is part of the reason why love cannot be wholly understood as a synchronic response to the subject's perspective at a given time on how things are. Like a lover, a love has a life. Just as understanding why a person is the way they are now requires us to look at their history, so fully understanding why a love is as it is now might require us to look at its history. This is one of the truths that is captured but misconstrued by the cognitivist relationship theory. On this view, it is at least in part the relationship itself, and the history of the love itself, not just the lover's knowledge of that relationship and that history, that explains the lover's present love. Moreover, unlike in the

[127]. Compare (Grau, 2010; Rorty, 1987).

case of attitudes and actions that are based on apparent reasons, this is not necessarily a kind of explanation that the subject has to be conscious of in order for the attitude to make sense to them, if 'making sense' of a state is a precondition of identifying with it.

6.6 Attraction, love and inanimate objects

We should of course acknowledge that love for people and 'love' for inanimate objects are not the same. Interpersonal love, because of the way it involves two parties, raises issues that are unique to it. Inanimate objects cannot love you back, cannot form expectations of you, and, in at least one ethically and psychologically significant sense, cannot be helped or harmed by you. There is no question of forming a shared life with an object or forming the kind of union of subjects that some[128] take to characterise love.

Nonetheless, some of the key features of love that I have discussed in this chapter do appear to be common to love of people and 'love' of mere objects—or places, or ideas, or sports, or styles of art, or pursuits, or whatever else we can become attached to in a way naturally expressed by the word 'love'.[129] The kind of partiality involved in love, wherein one takes a special interest in a person that cannot be explained merely in terms of that person's qualities, is just as much a feature of my love for the city of Bristol, for instance, or of the collector's 'love' for the mediocre bronze. The same goes for the particularity of these attitudes and hence the non-substitutability of their objects. The collector is drawn to *this* bronze. If it is true that, as I have argued, his attraction to the bronze is not rationally based on instantiating features of it that would be reasons for anyone in his circumstances to want to have it, this helps to explain the particularity that we naturally imagine characterising his attraction—the fact that he wants it and it alone, and not some alternative, not even an objectively superior one.

In this light, I think we can draw a very real connection between collector's attraction to the bronze and the actual process of falling in love. While the collector's desire is in a sense whimsical or fanciful, it has the potential to develop into something lasting and more significant, which it can only do if he acts on it. This would bring something of value into the collector's life; again, though, we cannot understand the desire as based on a recognition of this potential value.

It might be useful here to return once again to Kolodny's account and to consider what he says about the initiation and development of loving relationships. Love, on Kolodny's view, is justified by the value of a relationship, but the relationship has to be actual in order to make sense of love. You do not love someone because you *could* have a good relationship with them, but because you *do* have or have had one. Where do these relationships come from? One does not simply find oneself in a valuable relationship with a person, having no idea how one got there, and realise that, being in this situation, one ought to love this person.

Here is how Kolodny sketches the process of falling in love:

---

[128]. For example (Nozick, 1989; Solomon, 1981).

[129]. (Frankfurt, 2004) takes the commonality here very seriously—perhaps to the extent that he is insufficiently sensitive to the especially interpersonal issues in the case of interpersonal love.

> For one reason or another, you find yourself participating with Lisa, say, in activities of the kind that characterize established friendships, such as enjoying your leisure together, sharing a sense of humor, getting to know one another, exchanging confidences, providing assistance, and so on. Provided that nothing comes to light that would preclude a friendship with her, this pattern of interaction gradually gives rise to noninstrumental concern for Lisa … (Kolodny, 2003, p. 169)

… where 'noninstrumental concern' is, of course, the kind of concern involved in the kind of relationship that justifies love. So: first comes interaction, then interaction marked by concern and emotional vulnerability, then love. The question is what gets the interaction going in the first place.

This is not, in itself, a problem. You could get into a relationship by accident, or through an arranged marriage, for example. The relationship could end up being good and you could end up falling in love. There is always be a good deal of accident and contingency in our ending up with the people we end with. Nevertheless, it is often the can that people are more active than this in the development of their own relationships. One more active way of getting into a relationship—the eHarmony method, we might call it—is by being calculating. You know what kind of person you are and what kind of person you need to be with in order to have a good relationship. You find a person who meets the criteria and pursue a relationship with them. If they are interested too, hopefully things go well and you end up loving each other.

Clearly, though, there is another way in which one can play an active role that is not calculating: you feel attracted to another person, spend time with them, and a relationship grows. Normally, you want to spend time with someone not because you have assessed their merits as a potential partner, but because you feel drawn to them. In fact, people sometimes get into quite messy relationships well aware beforehand that things probably wouldn't work out, given what each of them is like. They do so because they have such strong feelings for each other. This might not be sensible, but it is not unintelligible. Relationships *can* get going without strong feelings of attraction, but a particularly natural way for them to get going is when people feel, and follow, such feelings.

Although Kolodny says less about attraction than about love, he does make some suggestive comments. He says that your being attracted to someone 'reflects that you do or would find engaging in certain activities with that person rewarding.' This, he says, is 'a reason to pursue those activities, and it is, in turn, a reason to want a relationship with that person in the context of which those activities might be pursued.' He suggests that this is particularly clear in the case of sexual attraction: you 'view [the other person's] charms as … making sex with him or her seem appealing' (Kolodny, 2003, p. 172). There may be some truth in this, but to echo the discussion of enjoyment in Chapter 3, it is not as straightforward as simply recognising and being motivated by reasons that would be reasons independently of one's responding to them. What kind of person you find appealing in this way really comes down to who you happen to find appealing, not to judgements about what features of a person would be reasons for anyone in your situation to 'engage with' in the

relevant way. In fact it is in this kind of case that this point seems to me to be the *most* obvious.

Kolodny, trying to articulate a picture of attraction that is consistent with a cognitivist framework and a universalistic conception of reasons, is led to say the following:

> Certain qualities cannot count as reasons for anyone to be attracted to a person. The weight of a person's kidneys or her social security number, for example, do nothing to render attraction to her intelligible. Nevertheless, we are fairly promiscuous about the qualities that we recognize as rendering attraction intelligible. Within these permissive limits, the reasons for attraction provided by these qualities are noninsistent reasons. Finding one *set of qualities* (within these limits) appealing is appropriate, but failing to find some other set (within these limits) appealing is not inappropriate. This judgment universalizes. It is open to everyone, but not required of everyone, to be attracted to this set of qualities. Moreover, one can have an insistent reason to be attracted to *a particular person*. Given that one is the *kind* of person who finds this *set of qualities* appealing, the fact that Jane has this set of qualities is an insistent reason to be attracted to Jane. This reason also universalizes. For everyone who finds such qualities appealing, the fact that Jane has such qualities is a reason to be attracted to her. (Kolodny, 2003, n. 38)

It seems to me a great strength of the view of attraction that I have been recommending that it does not commit us to saying such things. 'Having a type' is just a matter of being predisposed to be attracted to people with certain features. People of that kind tend, in general, to appeal to something in you that people who lack those features do not. Identifying someone's 'type' is a matter of inductive reasoning, of empirical generalisation. It does not have the normative significance that Kolodny is forced to accord to it.

6.7 The personal, the universal, and the intelligible

How, if not by subsumption under a universal principle, do we make the actions that are motivated by attraction or love intelligible? Explaining an action by saying that it was done out of a certain desire or out of love for a certain person does, like any explanation, bring something specific under something more general. We have a general understanding of how desire and love move us to action. The point I have tried to make is that this is not—not always—a matter of the *agent's* being moved to the action because they recognise that it falls under some general normative principle. When we act on worldly reasons (or apparent worldly reasons), our motivation does arguably have this structure. We move, in practical thought, from the general to the particular, and the same structure that motivates us also explains our actions.

Sometimes, though, what initiates movement in us is not the recognition of a universal reason, but rather something personal and particular, something that comes from within us. When a person acts on such a motivation, a proper appreciation of how their action made sense from their point of view as its agent is not fundamentally provided by seeing how that action would have made sense for anyone in the circumstances as the agent took them to be. It is founded instead on an appreciation of what it is like to desire or to care about something, to be moved by such desiring or caring into action, to be fulfilled or frustrated in

the actions that are so motivated. More generally and 'universally' we might bring to bear a conception of the place of such motives and actions in a good human life, but again, such a conception depends on our understanding of the character of these motivations themselves. In particular cases, gaining a deeper understanding of a person's motivation and action often requires not more and better metarepresentation—a more detailed picture of their picture of how the world is—but a richer understanding of the agent themselves, their history, their tastes, their character.

These idiosyncratic motivations are not responses to universal reasons and their force does not derive from their being backed up by universal reasons. However, neither do they simply fall outside the space of reasons. As we have seen, they can interact with universal reasons in a variety of ways. Reason can lead an agent to reject an idiosyncratic desire or to resist its force. It can consider whether acting on the desire would in the present case be on balance a good or a bad thing. In the central case, though, in which a person simply acts on such a desire, 'Reason' may turn a blind eye. The unruly part of the soul is allowed some free rein. When it is, we act for reasons, but the reasons for which we act are irreducibly particular, personal, and idiosyncratic.

References

Alston, W. P. (1983). What's wrong with immediate knowledge? *Synthese*, *55*(1), 73–95. https://doi.org/10.1007/BF00485374

Alvarez, M. (2010). *Kinds of Reasons*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199550005.001.0001/acprof-9780199550005

Andrews, K. (2003). Knowing mental states: The asymmetry of psychological prediction and explanation. In Q. Smith & A. Jokic (Eds.), *Consciousness: New Philosophical Perspectives* (pp. 201–219). Oxford: Oxford University Press.

Anscombe, G. E. M. (1963). *Intention* (2nd ed.). Ithaca, N.Y.: Cornell University Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Aristotle. (2011). *Aristotle's Nicomachean Ethics*. (R. C. Bartlett & S. D. Collins, Trans.). London: University of Chicago Press.

Arpaly, N. (2002). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Austin, J. L. (1950). Truth. *Proceedings of the Aristotelian Society Supplementary Volume*, *24*, 111–128.

Babb, M. (2016). The Essential Indexicality of Intentional Action. *The Philosophical Quarterly*, *66*(264), 439–457. https://doi.org/10.1093/pq/pqw023

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological Reasoning in Infancy. *Annual Review of Psychology*, *67*(1), 159–186. https://doi.org/10.1146/annurev-psych-010213-115033

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Båve, A. (2017). Self-Consciousness and Reductive Functionalism. *The Philosophical Quarterly*, *67*(266), 1–21. https://doi.org/10.1093/pq/pqw029

Bealer, G. (1997). Self-Consciousness. *The Philosophical Review*, *106*(1), 69–117. https://doi.org/10.2307/2998342

Blackburn, S. (1998). *Ruling Passions*. Oxford: Oxford University Press.

Blackburn, S. (2010). The Majesty of Reason. *Philosophy*, *85*(1), 5–27.

Block, N. (1980). Troubles with Functionalism. In *Readings in the Philosophy of Psychology, Volumes 1 and 2* (pp. 268–305). Cambridge, MA: Harvard University Press.

Block, N. (2007). *Consciousness, function, and representation*. Cambridge, Mass.; London: MIT. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=190963

Burge, T. (1979). Individualism and the Mental. *Midwest Studies In Philosophy*, *4*(1), 73–121. https://doi.org/10.1111/j.1475-4975.1979.tb00374.x

Burge, T. (1986). Individualism and Psychology. *The Philosophical Review*, *95*(1), 3–45. https://doi.org/10.2307/2185131

Burge, T. (2005). Disjunctivism and Perceptual Psychology. *Philosophical Topics*, *33*(1), 1–78.

Butler, J. (1729). *Fifteen Sermons Preached at the Rolls Chapel* (2nd ed.). London. Retrieved from http://find.galegroup.com/ecco/infomark.do?&source=gale&prodId=ECCO&userGroupName=ucl_ttda&tabID=T001&docId=CW423154603&type=multipage&contentSet=ECCOArticles&version=1.0&docLevel=FASCIMILE

Butterfill, S. A., & Apperly, I. A. (2013). How to Construct a Minimal Theory of Mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Byrne, A., & Logue, H. (2008). Either/Or. In A. Haddock & F. Macpherson (Eds.), *Disjunctivism: Perception, Action, Knowledge* (pp. 57–94). Oxford: Oxford University Press.

Cappelen, H., & Dever, J. (2013). *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person*. Oxford: Oxford University Press.

Chang, R. (Ed.). (1997a). *Incommensurability, incomparability, and practical reason*. Cambridge, Mass.: Harvard Univ. Press.

Chang, R. (1997b). *Incomparability and Practical Reason* (D.Phil. thesis). Oxford University, Oxford.

Chang, R. (Ed.). (1997c). Introduction. In *Incommensurability, incomparability, and practical reason* (pp. 1–34). Cambridge, Mass.: Harvard Univ. Press.

Chang, R. (2011). Can Desires Provide Reasons for Action? In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and value: themes from the moral philosophy of Joseph Raz* (Reprint, pp. 56–90). Oxford: Clarendon Press [u. a.].

Collins, A. W. (1997). The Psychological Reality of Reasons. *Ratio*, *10*(2), 108–123.

Comesaña, J., & McGrath, M. (2014). Having False Reasons. In C. Littlejohn & J. Turri (Eds.), *Epistemic Norms: New Essays on Action, Belief, and Assertion* (pp. 59–79). Oxford: Oxford University Press.

Copp, D., & Sobel, D. (2002). Desires, Motives, and Reasons: Scanlon's Rationalistic Moral Psychology. *Social Theory and Practice*, *28*(2), 243–276.

Cunningham, J. (2018). Knowledgeably Responding to Reasons. *Erkenntnis*. https://doi.org/10.1007/s10670-018-0043-3

Dancy, J. (2000). *Practical Reality*. Oxford: Oxford University Press.

Davidson, D. (1980a). Actions, Reasons, and Causes. In D. Davidson, *Essays on Actions and Events* (pp. 3–19). Oxford: Oxford University Press.

Davidson, D. (1980b). *Essays on Actions and Events*. Oxford: Oxford University Press.

Davidson, D. (1980c). How is the Weakness of the Will Possible? In *Essays on Actions and Events* (pp. 25–42). Retrieved from https://ci.nii.ac.jp/naid/10026654646/

Davidson, D. (1991). Three Varieties of Knowledge. *Royal Institute of Philosophy Supplement*, *30*, 153–166.

Davidson, D. (2001). *Inquiries into Truth and Interpretation* (2nd ed.). Oxford: Oxford University Press.

Davis, W. A. (1986). Two Senses of Desire. In *The Ways of Desire* (pp. 181–196). Precedent.

Döring, S. A., & Eker, B. (2017). Desires without Guises. In J. A. Deonna & F. Lauria (Eds.), *The nature of desire* (pp. 79–103). New York, NY: Oxford University Press.

Dretske, F. (1994). If You Can't Make One, You Don't Know How It Works. *Midwest Studies In Philosophy*, *19*(1), 468–482. https://doi.org/10.1111/j.1475-4975.1994.tb00299.x

Eilan, N. (2011). Experiential Objectivity. In J. Roessler, H. Lerman, & N. Eilan (Eds.), *Experiential Objectivity* (pp. 51–67). Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199692040.001.0001/acprof-9780199692040-chapter-4

Engel, P. (2013). Doxastic Correctness. *Aristotelian Society Supplementary Volume*, *87*(1), 199–216. https://doi.org/10.1111/j.1467-8349.2013.00226.x

Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.

Fodor, J. A. (1968). *Psychological Explanation*. New York: Random House.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97–115. https://doi.org/10.1007/BF00485230

Fodor, J. A. (1991). A Modal Argument for Narrow Content. *The Journal of Philosophy*, *88*(1), 5–26. https://doi.org/10.2307/2027084

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5–20.

Frankfurt, H. G. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Frankfurt, H. G. (2004). *The Reasons of Love*. Oxford: Princeton University Press.

Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gibbons, J. (2001). Knowledge in Action. *Philosophy and Phenomenological Research*, *62*(3), 579–600. https://doi.org/10.1111/j.1933-1592.2001.tb00075.x

Goldman, A. H. (2009). *Reasons from Within*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199576906.001.0001/acprof-9780199576906?result=1&rskey=Vkke7i&q=reasons%20from%20within

Goldman, A. I. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy*, *73*(20), 771–791. https://doi.org/10.2307/2025679

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, *7*(1–2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Grau, C. (2010). Love and History. *The Southern Journal of Philosophy*, *48*(3), 246–271. https://doi.org/10.1111/j.2041-6962.2010.00030.x

Hampshire, S. (1999, April 22). The Reason Why Not. *The New York Review of Books*. Retrieved from https://www.nybooks.com/articles/1999/04/22/the-reason-why-not/

Harman, G. H. (1964). How Belief is Based on Inference. *The Journal of Philosophy*, *61*(12), 353–359.

Harrigan, K., Hacquard, V., & Lidz, J. (2018). Three-Year-Olds' Understanding of Desire Reports Is Robust to Conflict. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00119

Heal, J. (2003). *Mind, Reason and Imagination*. Cambridge: Cambridge University Press.

Heathwood, C. (2006). Desire Satisfactionism and Hedonism. *Philosophical Studies*, *128*(3), 539–563. https://doi.org/10.1007/s11098-004-7817-y

Heuer, U. (2004). Reasons for Actions and Desires. *Philosophical Studies*, *121*(1), 43–63.

Hinton, J. M. (1973). *Experiences: an inquiry into some ambiguities*. Oxford: Clarendon Press.

Holloway, P. (1966). *Love and Desire*. Los Angeles: Capitol Records, Inc.

Hornsby, J. (2008). A Disjunctive Conception of Acting for Reasons. In A. Haddock & F. Macpherson (Eds.), *Disjunctivism: Perception, Action, Knowledge* (pp. 244–261). Oxford: Oxford University Press.

Hughes, N. (2014). Is Knowledge the Ability to ϕ for the Reason That P? *Episteme*, *11*(4), 457–462. https://doi.org/10.1017/epi.2014.16

Hyman, J. (1999). How Knowledge Works. *Philosophical Quarterly*, *50*(197), 433–451.

Hyman, J. (2006). Knowledge and Evidence. *Mind*, *115*(460), 891–916.

Hyman, J. (2010). The Road to Larissa. *Ratio*, *23*(4), 393–414.

Hyman, J. (2011). Acting for Reasons: Reply to Dancy. *Frontiers of Philosophy in China*, *6*(3), 358–368.

Hyman, J. (2015). *Action, knowledge, and will* (First Edition). Oxford, United Kingdom ; New York, NY: Oxford University Press.

Hyman, J. (2017). Knowledge and Belief. *Aristotelian Society Supplementary Volume*, *91*(1), 267–288. https://doi.org/10.1093/arisup/akx005

Kolodny, N. (2003). Love as Valuing a Relationship. *The Philosophical Review*, *112*(2), 135–189.

Kolodny, N. (2005). Why Be Rational? *Mind*, *114*(455), 509–563. https://doi.org/10.1093/mind/fzi509

Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

Lavin, D. (2011). Problems of Intellectualism: Raz on Reason and Its Objects. *Jurisprudence*, *2*, 367–378.

Leslie, A. M. (1987). Pretense and Representation: The Origins of 'Theory of Mind'. *Psychological Review*, *94*(4), 412–426.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249–258. https://doi.org/10.1080/00048407212341301

Lewis, D. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *63*, 113–137.

Littlejohn, C. (2012). *Justification and the truth-connection*. Cambridge: Cambridge University Press.

Littlejohn, C. (2014). Fake Barns and False Dilemmas. *Episteme*, *11*(4), 369–389. https://doi.org/10.1017/epi.2014.24

Littlejohn, C. (2015). Knowledge and Awareness. *Analysis*, *75*(4), 596–603. https://doi.org/10.1093/analys/anv051

Loar, B. (1981). *Mind and meaning*. Cambridge [usw.]: Cambridge Univ. Pr.

Locke, D. (2015). Knowledge, Explanation, and Motivating Reasons. *American Philosophical Quarterly*, *52*(3), 215–232.

MacFarquhar, L. (2011, 05). How To Be Good. *The New Yorker*. Retrieved from https://www.newyorker.com/magazine/2011/09/05/how-to-be-good

Martin, M. G. F. (1999). *Desire in Time*. Unpublished manuscript.

Martin, M. G. F. (2004). The Limits of Self-Awareness. *Philosophical Studies*, *120*(1), 37–89. https://doi.org/10.1023/B:PHIL.0000033751.66949.97

Martin, M. G. F. (2006). On Being Alienated. In T. S. Gendler & J. Hawthorne (Eds.), *Perceptual Experience*. Oxford: Oxford University Press.

Mayr, E. (2011). *Understanding Human Agency*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199606214.001.0001/acprof-9780199606214

McCullagh, M. (2000). Functionalism and Self-Consciousness. *Mind & Language*, *15*(5), 481–499. https://doi.org/10.1111/1468-0017.00146

McDowell, J. (1985). Functionalism and Anomalous Monism. In E. LePore & B. P. McLaughlin (Eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (pp. 387–398). Oxford: Basil Blackwell.

McDowell, J. (1994). Knowledge by Hearsay. In B. K. Matilal & A. Chakrabarti (Eds.), *Knowing from Words: Western and Indian Philosophical Analysis of Understanding and Testimony* (Vol. 230, pp. 195–224). Dordrecht: Kluwer.

McDowell, J. (1995). Knowledge and the Internal. *Philosophy and Phenomenological Research*, *55*(4), 877–893.

McDowell, J. (1998a). Criteria, Defeasibility, and Knowledge. In *Meaning, Knowledge, and Reality* (pp. 369–394). London: Harvard University Press.

McDowell, J. (1998b). Functionalism and Anomalous Monism. In *Mind, Value, and Reality* (pp. 325–340). London: Harvard University Press.

McDowell, J. (2013). Acting in the Light of a Fact. In D. Bakhurst, M. O. Little, & B. Hooker (Eds.), *Thinking About Reasons: Themes From the Philosophy of Jonathan Dancy* (pp. 13–28). Oxford: Oxford University Press.

McGinn, C. (1977). Charity, interpretation, and belief. *Journal of Philosophy*, *74*, 521–535.

Millgram, E. (1997). *Practical Induction*. London: Harvard University Press.

Millgram, E., & Thagard, P. (1996). Deliberative coherence. *Synthese*, *108*(1), 63–88. https://doi.org/10.1007/BF00414005

Moore, C., Jarrold, C., Russell, J., Lumb, A., Sapp, F., & MacCallum, F. (1995). Conflicting desire and the child's theory of mind. *Cognitive Development*, *10*(4), 467–482. https://doi.org/10.1016/0885-2014(95)90023-3

Nagel, T. (1978). *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

Nozick, R. (1989). Love's Bond. In *The Examined Life: Philosophical Meditations* (pp. 68–86). New York: Simon & Schuster.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Parfit, D. (2011). *On what matters* (Vol. 1). Oxford; New York: Oxford University Press.

Peacocke, C. (1986). *Thoughts: An Essay on Content*. Oxford: Basil Blackwell.

Peacocke, C. (1993). Externalist Explanation. *Proceedings of the Aristotelian Society*, *93*, 203–230.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA, US: The MIT Press.

Perner, J., Brandl, J. L., & Garnham, A. (2003). What is a Perspective Problem? Developmental Issues in Belief Ascription and Dual Identity. *Facta Philosophica*, *5*, 355–378.

Perner, J., Priewasser, B., & Roessler, J. (2018). The Practical Other: Teleology and its Development. *Interdisciplinary Science Reviews*, *43*(2).

Perner, J., Rendl, B., & Garnham, A. (2007). Objects of Desire, Thought, and Reality: Problems of Anchoring Discourse Referents in Development. *Mind & Language*, *22*(5), 475–513. https://doi.org/10.1111/j.1468-0017.2007.00317.x

Perner, J., Zauner, P., & Sprung, M. (2005). What Does "That" Have to Do with Point of View? In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 220–244). New York; Oxford: Oxford University Press.

Perry, J. (1979). The Problem of the Essential Indexical. *Noûs*, *13*(1), 3–21. https://doi.org/10.2307/2214792

Plato. (1952). *Plato's Phaedrus*. (R. Hackforth, Trans.). Cambridge: Cambridge University Press.

Pollock, J. (1986). *Contemporary Theories of Knowledge*. Savage, MD: Rowman and Littlefield.

Prior, A. N. (1971). *Objects of Thought*. (P. T. Geach & A. J. P. Kenny, Eds.). Oxford: Oxford University Press.

Pryor, J. (2000). The Skeptic and the Dogmatist. *Noûs*, *34*(4), 517–549. https://doi.org/10.1111/0029-4624.00277

Putnam, H. (1975). The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.

Quinn, W. (1994). Putting Rationality in Its Place. In *Morality and Action* (pp. 228–255). Cambridge: Cambridge University Press.

Railton, P. (2012). That Obscure Object, Desire. *Proceedings and Addresses of the American Philosophical Association*, *86*(2), 22–46.

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). "This way!", "No! That way!"—3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, *22*(1), 47–68. https://doi.org/10.1016/j.cogdev.2006.08.002

Raz, J. (1986). *The Morality of Freedom*. Oxford: Clarendon Press.

Raz, J. (1997). When We Are Ourselves: The Active and the Passive. *Aristotelian Society Supplementary Volume*, *71*(1), 211–228. https://doi.org/10.1111/1467-8349.00027

Raz, J. (1999). *Practical Reason and Norms* (rev. ed.). Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198268345.001.0001/acprof-9780198268345

Raz, J. (2000). *Engaging Reason: On the Theory of Value and Action*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/0199248001.001.0001/acprof-9780199248001

Raz, J. (2003). *The Practice of Value* (Rev. ed.). Oxford: Oxford University Press.

Repacholi, B. M., & Gopnik, A. (1997). Early Reasoning About Desires: Evidence From 14- and 18-Month-Olds. [Miscellaneous Article]. *Developmental Psychology*, *33*(1), 12–21.

Rescher, N. (2009). Choice Without Preference: A Study of the History and of the Logic of the Problem of "Buridan's Ass". *Kant-Studien*, *51*(1–4), 142–175. https://doi.org/10.1515/kant.1960.51.1-4.142

Rey, G. (2007). Resisting normativism in psychology. In B. P. McLaughlin & J. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind* (pp. 69–84). Oxford: Blackwell.

Robinson, H. (1985). The General Form of the Argument for Berkeleian Idealism. In J. Foster & H. Robinson (Eds.), *Essays on Berkeley: A Tercentennial Celebration*. Oxford: Clarendon Press.

Robinson, H. (1994). *Perception*. London: Routledge.

Roessler, J. (2014). Reason Explanation and the Second-Person Perspective. *Philosophical Explorations*, *17*(3), 346–357.

Roessler, J., & Perner, J. (2013). Teleology: Belief as Perspective. In S. Baron-Cohen, H. Tager-Flusberg, & M. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 35–50).

Rorty, A. O. (1987). The Historicity of Psychological Attitudes: Love Is Not Love Which Alters Not When It Alteration Finds. *Midwest Studies In Philosophy*, *10*(1), 399–412. https://doi.org/10.1111/j.1475-4975.1987.tb00548.x

Scanlon, T. (1998). *What We Owe to Each Other* (Nachdr.). London: Harvard University Press.

Schapiro, T. (2014). What are Theories of Desire Theories of? *Analytic Philosophy*, *55*(2), 131–150. https://doi.org/10.1111/phib.12043

Scheffler, S. (2010). *Equality and tradition: questions of value in moral and political theory*. New York: Oxford University Press.

Schroeder, M. (2007). *Slaves of the Passions*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199299508.001.0001/acprof-9780199299508-miscMatter-4

Schroeder, T. (2004). *Three faces of desire*. Oxford; New York: Oxford University Press.

Schueler, G. F. (1995). *Desire: its role in practical reason and the explanation of action*. London: MIT Press.

Setiya, K. (2014). Love and the Value of a Life. *The Philosophical Review*, *123*(3), 251–280. https://doi.org/10.1215/00318108-2683522

Shah, N. (2003). How Truth Governs Belief. *The Philosophical Review*, *112*(4), 447–482.

Shoemaker, S. S. (1968). Self-Reference and Self-Awareness. *The Journal of Philosophy*, *65*(19), 555–567. https://doi.org/10.2307/2024121

Siegel, S. (2004). Indiscriminability and the Phenomenal. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *120*(1/3), 91–112.

Smith, M. (1987). The Humean Theory of Motivation. *Mind*, *96*(381), 36–61.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Snowdon, P. (2005). The Formulation of Disjunctivism: A Response to Fish. *Proceedings of the Aristotelian Society*, *105*(1), 129–141. https://doi.org/10.1111/j.0066-7373.2004.00106.x

Solomon, R. C. (1981). *Love: Emotion, Myth, and Metaphor*. New York: Anchor Press.

Stalnaker, R. (1989). On What's In the Head. *Philosophical Perspectives*, *3*, 287–316. https://doi.org/10.2307/2214271

Stalnaker, R. (1990). Narrow Content. In C. A. Anderson & J. Owens (Eds.), *Propositional Attitudes: The Role of Content in Logic, Language, and Mind* (pp. 131–145). Stanford: CSLI Publications.

Stampe, D. W. (1987). The Authority of Desire. *The Philosophical Review*, *96*(3), 335–381. https://doi.org/10.2307/2185225

Stich, S. P. (1985). *From folk psychology to cognitive science: the case against belief*. Cambridge, MA: MIT Press.

Stocker, M. (2011). Raz on the Intelligibility of Bad Acts. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and value: themes from the moral philosophy of Joseph Raz* (Reprint, pp. 303–332). Oxford: Clarendon Press [u. a.].

Strawson, P. F. (1950). Truth. *Proceedings of the Aristotelian Society Supplementary Volume*, *24*, 129–156.

Stroud, B. (2000). *The Quest for Reality*. Oxford: Oxford University Press.

Sturgeon, S. (1998). Visual Experience. *Proceedings of the Aristotelian Society*, *98*, 179–200.

Sylvan, K. (2015). What apparent reasons appear to be. *Philosophical Studies*, *172*(3), 587–606. https://doi.org/10.1007/s11098-014-0320-1

Taylor, G. (1975). Love. *Proceedings of the Aristotelian Society*, *76*, 147–164.

Tooley, M. (2001). Functional Concepts, Referentially Opaque Contexts, Causal Relations, and the Definition of Theoretical Terms. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *105*(3), 251–279.

Ullmann-Margalit, E., & Morgenbesser, S. (1977). Picking and Choosing. *Social Research*, *44*(4), 757–785.

Unger, P. K. (1975). *Ignorance: a case for scepticism*. Oxford: Clarendon Press.

Vogler, C. (2002). *Reasonably Vicious*. London: Harvard University Press.

Watson, G. (1975). Free Agency. *The Journal of Philosophy*, *72*(8), 205–220. https://doi.org/10.2307/2024703

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, *16*(s16), 267–97.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Estes, D. (1986). Early Understanding of Mental Entities: A Reexamination of Childhood Realism. *Child Development*, *57*(4), 910–923. https://doi.org/10.2307/1130367

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275. https://doi.org/10.1016/0010-0277(90)90024-E

Williams, B. (1981a). Internal and External Reasons. In *Moral luck: philosophical papers, 1973-1980* (pp. 101–113). Cambridge: Cambridge University Press.

Williams, B. (1981b). Persons, character and morality. In *Moral Luck* (pp. 1–19). London: Cambridge University Press.

Williams, B. (1985). *Ethics and the Limits of Philosophy*. London: Fontana Press.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Wollheim, R. (1984). *The thread of life*. New Haven: Yale University Press.

Wollheim, R. (1999). *On the emotions*. New Haven, CT: Yale University Press.

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

Yablo, S. (2003). Causal Relevance. *Philosophical Issues*, *13*(1), 316–329. https://doi.org/10.1111/1533-6077.00016

Yalowitz, S. (1997). Rationality and the Argument for Anomalous Monism. *Philosophical Studies*, *87*(3), 235–258. https://doi.org/10.1023/A:1004200832034

Yalowitz, S. (2014). Anomalous Monism. Retrieved from https://plato.stanford.edu/archives/win2014/entries/anomalous-monism/

Yao, V. (forthcoming). The Undesirable & The Adesirable. *Philosophy and Phenomenological Research*, *Early View*. https://doi.org/10.1111/phpr.12475

Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. *British Journal of Developmental Psychology*, *2*(1), 73–81. https://doi.org/10.1111/j.2044-835X.1984.tb00536.x

Yuill, N., Perner, J., Pearson, A., Peerbhoy, D., & Ende, J. van den. (1996). Children's changing understanding of wicked desires: From objective to subjective and moral. *British Journal of Developmental Psychology*, *14*(4), 457–475. https://doi.org/10.1111/j.2044-835X.1996.tb00718.x