Table of Contents

Abstract 4
Declaration 5
Copyright Statement 5
Acknowledgements 6

Introduction 7

Chapter One: Sentiments, Emotions, and Caring 15
1. Introduction 15
2. Emotions and caring 18
   2.1 Assumptions about emotion 18
   2.2 The connection with cares 22
3. The Simple View of cares 25
4. Antirealist interpretations of the Simple View 31
5. Counterexamples to antirealism 35
   5.1 Patterns without cares 35
   5.2 Cares without patterns 41
6. The Simple View without patterns 44
7. Conclusion 46

Chapter Two: The Nature of Dispositions 47
1. Introduction 47
2. The problem with dispositions 49
3. Preliminary distinctions 50
   3.1 Realism and antirealism about kinds of properties 51
   3.2 Covert/conventional and overt/canonical disposition ascriptions 53
4. The simple conditional analysis 54
5. Counterexamples to SCA: Finks, masks, and mimics 56
6. Lewis’ reformed conditional analysis 60
7. Saving the analysis 66
   7.1 Getting specific: Modifying the analysans 67
   7.2 Getting specific: Modifying the analysandum 70
8. Are dispositions an intrinsic matter? 75
9. Intrinsic finks, masks and mimics 79
10. Towards pluralism 83
11. The case for intrinsic psychological dispositions 85
12. The case for extrinsic psychological dispositions 87
13. Conclusion 89

Chapter Three: Emotional Dispositions 90
1. Introduction 90
2. Psychological and non-psychological explanations 92
3. Variety of emotional dispositions 93
4. Wollheim’s defence of realism about psychological dispositions 96
Chapter Four: Caring, Attitudes, and Character

1. The realist-dispositional account of cares
2. The ethical significance of the mental
3. Psychological dispositions as attitudes
4. What kind of psychological disposition is a care?
   4.1 The centrality of caring
   4.2 Caring as attitude
   4.3 Filling in the details of the realist-dispositional account
5. Caring and character

Chapter Five: The Reality of Character: A Reply to Situationism

1. Introduction
2. From thick concepts to empirical predictions
3. Varieties of situationism and implications for virtue ethics
4. Character traits reduction
5. Character traits as ‘tendencies’ (realistically construed)
6. A non-reductive, non-tendential account of character traits
7. Virtue ethics and weak situationism
8. Conclusion

Chapter Six: Sentiments and Reasons

1. Introduction
2. Rationalism vs. antirationalism about caring and love
3. Prima facie reasons to be at least local antirationalists
   3.1 Projection
   3.2 Attachment
   3.3 Does the account generalise? Two observations
4. Love for particulars
5. Two puzzles about romantic love
6. Options for the rationalist
7. Kolodny’s relationship account
8. Intrinsic properties contextualised
9. Synchronic and diachronic justification
10. Conclusion: The possibility of refined emotions

Appendix: A Reply to Deonna and Teroni (2012)

Bibliography

Word count: 75 624
Abstract

Contemporary emotion research typically takes the phenomenon of emotion to be exhausted by a class of mental events that are intentional, conscious, and related to certain sorts of behaviour. Moreover, other affective phenomena, such as moods, are also considered to be relatively short-term, episodic, or occurrent states of the subject undergoing them. Emotions, and other putative emotional phenomena that common-sense takes as long-lasting, non-episodic, or dispositional are things that both philosophers and scientists sometimes recognise, but that are relatively neglected in comparison to emotional episodes. This thesis aims at showing that this neglect is unjustified. I will argue that there is a class of entities, 'sentiments'—broadly characterised as dispositions to undergo emotional episodes—that (1) are irreducible to emotional episodes or collections thereof and (2) have properties that make them a suitable target of study by the emotion researcher.

In the first chapter, I argue that an analysis of caring (and related phenomena, such as love) as a pattern of emotional episodes, while more plausible than alternative, non-emotional accounts, faces a number of counterexamples that motivate the search for an account of caring as related in a certain way to emotions but as irreducible to them. I argue that a dispositional account, according to which dispositions are conceived as distinct from their manifestations, is an account for which a strong case can be made.

The second chapter is dedicated to defending a modest form of realism about dispositions in general and psychological dispositions in particular. According to realism, dispositions are genuine properties that, although perhaps reducible to non-dispositional properties, cannot be re-described in terms of events (including behaviour) only.

In the third chapter, I show in what ways emotional dispositions (or sentiments) can positively contribute to the explanation of the occurrence (or non-occurrence) of emotional episodes.

In the fourth chapter, I argue that caring, understood as a species of sentiment, is not to be construed as a mere disposition to produce certain events; rather, we should allow that certain dispositions are genuinely mental or psychological. Assuming realism about the mental, I argue that some dispositions are mental in a way that others (such as fragility) are not. I suggest that being intentional is the property that makes psychological dispositions genuinely mental. I end the chapter by drawing a connection between caring and the notion of character. On my view, caring is at least a necessary ingredient of certain character traits, in particular the virtues.

In chapter five, I tackle a recent form of empirically informed scepticism about character and argue, on the basis of general considerations about psychological dispositions, that the sceptic’s case is not as strong as she makes out.

Finally, in chapter six, I argue that at least certain forms of sentiment, for example romantic love, can be genuinely supported by reasons, thereby suggesting a way they can contribute to the value of our lives.

Overall, the aim of this thesis is to establish the respectability of sentiments in a sophisticated account of the mind.
Declaration

I declare that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time.

The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “intellectual property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in the thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.campus.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

Many people have contributed, always in their own way, to the shape this thesis has taken over the years.

First, I’d like to thank the people with whom I discussed (and, sometimes, on whom I tested) what became the ideas presented in this work and who have provided very useful input. In particular, I’d like to thank Tom Bates, Phil Brown, Florian Cova, Sean Crawford, Luca Csepely-Knorr, Julien Deonna, Ana Sofia Ganho, Federico Lauria, Phil Letts, Clare Mac Cumhaill, Alberto Masala, Tony Milligan, Michiru Nagatsu, Eric Schwitzgebel, Nick Smyth, Fabrice Teroni, Lucy Tomlinson, and various audiences in Manchester, Paris, Lille, and Geneva. Special thanks go to Moritz Mueller for all the valuable discussions we’ve had on emotion since Day 1 of the PhD.

I’ve greatly benefitted from the comments of various people on parts of the manuscript. Thanks go to Christophe Al-Saleh, Tom Bates, Stefano Cossara, Chris Daly, Alexandre Marcellesi, Alberto Masala, Moritz Mueller, and Joel Smith. In particular, thanks to Michael Scott and Florian Cova for giving me detailed feedback on the entire manuscript.

I’d also like to thank Sophie Grenat for her patience and support and my parents and sister for always believing that I could go so far as getting a PhD. Thanks also for the inspiration I had when writing the material on caring and love.

My deepest gratitude goes to three people. First, my supervisor, Michael Scott, has been a wonderful teacher throughout the PhD, and I’ll always be thankful for his making me realize how philosophy—good philosophy, that is—should be done. I hope that this thesis is not too far from satisfying your standards, Michael.

Second, I’d like to thank Joel Smith for his support over the years. I cannot think of a time when Joel was not there for me, and I owe him much of the confidence I had to finish this work. Thanks for believing in me, Joel.

Last, but not least, I’d like to thank my former supervisor, Peter Goldie. No words can describe how lucky I feel to have met this man. It wouldn’t be an overstatement to say that he is the main reason I got into emotion research in the first place. His unexpected death, in October 2011, was a great loss for both philosophy and his students. He had such an inspiration on my work. I’ll never forget the conversation we had, at the Blackwell coffee place near the Manchester philosophy department (sometime in September 2010), where Peter first suggested to me that I look at the literature on dispositions and expressed his ideas on the possibility of masking cases in love. It was with this discussion in mind that most of the thesis was written. I hope Peter would have liked the result.
Introduction

As typically noted in the literature on emotion, two very broad kinds of emotions can be distinguished: on the one hand, we have ‘occurrent’, ‘episodic’, ‘short-term’ emotions such as fear at an approaching animal or anger at an offence; on the other hand, we have ‘dispositional’, ‘long-term’ ones such as one’s love for one’s partner or one’s long-lasting jealousy towards a competitor. Although the term ‘emotion’ is these days used to refer primarily to the former, it can also refer to the latter (Goldie, 2000). Following custom, I will reserve the term ‘emotion’ (or ‘emotional episode’) to refer to occurrent emotions and use instead the term ‘sentiment’ to refer to dispositional, long-term ones.

Surprisingly, most writers on the topic, after having made the distinction, go on to focus exclusively on emotional episodes. As a result, despite (arguably) a few notable exceptions (Wollheim, 1999, Deonna & Teroni, 2009), nothing like a systematic investigation into the nature of sentiments has been conducted in recent writings. If most writers are aware of the distinction, however, it is difficult to see this neglect as purely accidental and not motivated by a deeper worry about the category of sentiments. There indeed seems to be in the literature an assumption that emotions are more important in some way (e.g., Tappolet, 2002). A plausible explanation of this neglect is that sentiments are usually not thought to be distinct from emotions; emotions are more basic, and sentiments can be analysed in terms of them. For instance, Marcel’s love for Albertine may be nothing more than a complex entity essentially involving emotional episodes (plus perhaps some other kinds of mental events, such as thoughts and desires). For Paul Ekman, (what I call) a sentiment is nothing more than a summation of emotional episodes, as the following quotation suggests: “I believe those who claim emotions endure for much longer time periods are summarizing what is actually a series of briefer emotion episodes” (1994, 16, quoted in Goldie, 2000, 104). If this is right, the distinction between sentiments and emotions turns out to be a distinction between a whole and its parts or constituents. As a consequence, unless a sentiment has interesting properties of
its own, which for certain writers does not appear to be the case,¹ there seems to be no point in studying it in addition to the discrete emotional episodes that constitute it.

There are two main ways to resist this conclusion. First, one could argue that sentiments, as complexes or patterns of emotional episodes, do have a role to play over and above the one played by the mental events that constitute them. For instance, sentiments could make discrete mental events intelligible to the subject by placing them in a broader context or ‘narrative’ (Goldie, 2000, Shoemaker, 2003), or they could contribute to their justification in some way (Helm, 2001). Notice though that, being on such views collections of emotions, sentiments seem dispensable when it comes to providing a full explanation of why the emotions occur the way they do. To quote Richard Wollheim (criticising a similar view about dispositions in general), a sentiment is explanatory “only in the limited sense that it takes away the oddity that the single happening might otherwise have.” (1999, 5) The second option, therefore, is to argue that sentiments and emotions should be conceived as distinct entities that may nonetheless entertain a close relationship with each other. In this thesis, I pursue the second option.

One way to grant more importance to sentiments than they have on alternative, widely accepted accounts is to argue that they play an explanatory role that cannot be simply reduced to the explanatory role played by collections of discrete emotional episodes. If a sentiment were a mere collection of mental events, anything that a sentiment would do would be something that the collection of mental events does; in other words, it would have no explanatory role to play over and above the mosaic of emotional episodes that we feel over time, except perhaps in the limited sense pointed out by Wollheim.

It is instructive to look at a debate found in another context which bears some resemblances with the debate under discussion. Sentiments, on the view under consideration, are conceived as patterns of emotional episodes in a similar way

¹I take most writers’ silence on the matter to imply at least the tacit assumption that this is true.
laws of nature are construed by some Humeans. On a simple version of empiricism about laws of nature—called the ‘regularity’ theory—laws of nature should not be conceived as unobservable entities governing the events that we observe and experience, but as at best regularities in these events. Laws of nature, on such a view, are ways we have found of making intelligible disparate physical events by summarising them into single statements. If we were asked ‘where’ to locate laws of nature, we would probably be answered one of two things: either laws of nature are nothing more than regularities or patterns of physical events, or as summary statements capturing regularities in physical events, they are, if anywhere, ‘in our heads’, a simple product of our capacities of abstraction; either way, they would hardly be the kind of thing that can govern a whole universe. And one might claim that this is precisely what is problematic about the view (e.g., Carroll, 1994). On an ordinary conception of them, it seems, laws of nature are indeed supposed to do something, namely govern the physical events and regularities we can observe, and thereby explain them. It is, on such a conception, because laws of nature are a certain way that the physical events that we can observe (in addition to those we don’t observe, cannot observe and will never observe) occur in a certain way. Laws of nature, therefore, are distinct from both our minds and the patterns of physical events that occur in the world. So, even if the empiricist takes herself to believe in laws of nature, it would only be in an attenuated sense of the notion (Mumford, 2004, 32).

The question, of course, is why we should believe in laws of nature as things other than regularities between events. One answer is that, given that laws of nature obviously exist, and that the concept of a law of nature is not the concept of a regularity in events (or anything in the vicinity of the Humean’s worldview) but of something distinct from it, so laws of nature should not be reduced to

---

2 On a Ramsey-Lewis’ ‘best systems account’, “a contingent generalisation is a law of nature if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength” (Lewis, 1973, 73). A natural thought is that, since on such view laws of nature depend on human standards, those of simplicity and strength, they are mind-dependent entities. If that’s right, the best systems theorist is committed to the claim that laws of nature would not exist if minds did not exist. In a similar vein, if we take ‘summary statements’ to be the products of our minds, then we are committed to laws of nature being the products of our minds.
regularities in events. I take it that this sort of argument would move virtually no Humean. First, the Humean could easily claim that if the realist is right about the concept of a law of nature, then it is hardly obvious that laws of nature thus understood exist; the realist begs the question against the Humean by helping herself with an assumption—that laws-of-nature-as-the-realist-thinks-of-them exist—that the Humean is happy to reject. Second, the Humean could deny that it is a conceptual truth that laws of nature are over and above regularities in events, that they must ‘govern’ them, or whatnot. And if it is indeed true that laws of nature as the folk think of them are to be construed along realist lines, a little conceptual revision may be in order if we don’t need to postulate them but still would like to use the notion. It may be handy to keep our concept of a law of nature as it is, but if we have no good reason to believe in laws of nature thus understood, we had better do something about it (say, by revising or eliminating it). What we must ask ourselves, I contend, is whether there are good reasons to posit an entity—such as a law of nature—to explain the regularities (or some subset thereof) that we observe in the world, and what such an entity must be like.

The main motivation for the postulation of an extra kind of entity, although I wouldn’t call it an argument, is a deeply felt need to find appropriate explanations for certain facts of the world. One such fact is that the world includes events and regularities between events that do not seem purely accidental. It is no accident, it seems, that when you drop a ball on earth, it falls down and doesn’t go up, that no human being has managed to live longer than one thousand years, or that vases shatter when dropped and do not turn into daisies. As Galen Strawson puts it, “It may seem extraordinary that anyone should ever have held the view that there is definitely no ‘because’ in nature; that there is definitely nothing about the world in virtue of which it is regular.” (1989, 86) Whether or not this is ultimately a reasonable attitude, the desire to find adequate explanations for the relevant regularities has been at the basis of the postulation of chiefly non-Humean entities such as certain higher-order

3 See Beebee, 2000, for a forceful, and to my mind successful, attack on the claim that it is a conceptual truth that laws of nature are the kind of thing to ‘govern’.
4 Beebee (2006, 521) makes this claim regarding the anti-Humean account of causation.
universals (Armstrong, 1983) and dispositions or causal powers that are intrinsic to objects (Ellis, 2001, Mumford, 2004, Strawson, 1989) in terms of which laws of nature may in turn be analysed. Each of these entities has been claimed to provide the metaphysical basis for the various events that we can observe, and the regularities that they exhibit. It is in virtue of entities of these sorts, the anti-Humean tells us, that the Humean’s summary statements can be true and that we can distinguish events that are accidental from those that aren’t. So, whereas the Humean provides a description of the various events found in the world, the anti-Humean provides an explanation of them, an explanation appealing to entities distinct from the regularities. The success of the anti-Humean’s agenda depends in great part on the success of the particular explanations that she gives.

The motivation at the basis of this thesis is also a deeply felt need to explain, not the occurrence of physical events and of patterns or regularities between them, but the occurrence of a certain class of mental events, namely emotions. On the face of it, the fact that we experience them at times is perfectly unproblematic. Yet we can legitimately ask why we feel the emotions we feel at the moments we feel them, why other people may not feel them in similar circumstances, and why we may feel differently over time in objectively very similar situations. Moved by the anti-Humean intuition that explanations of events that appeal to other events are incomplete, I believe such questions should be answered in a similar way the issue about regularities in nature is answered by the anti-Humean: by positing a kind of entity that explains the relevant events without being reducible to them. I believe that in the case of emotions, this role is performed by sentiments.

We can now clearly see the similarity between the realism/reductionism debate about laws of nature and the debate (as yet to be begun) between the ‘pattern’ theorist about sentiments—we will see, a reductive account—and the one I will

---

5 Mumford (2004) takes powers to explain regularities in physical events, but thinks that laws of nature as ordinarily conceived cannot be identified with them. He argues that, since powers do all the explanatory work we wanted laws of nature to do, laws of nature become redundant and should in turn be eliminated. Ellis (2001) disagrees and takes certain dispositions to constitute laws of nature. Cartwright (1980) takes laws of nature to be descriptions of the causal powers of objects.
call the ‘realist’. If we are to be realists about sentiments, we may start by asking ourselves what the ordinary conception of them is, whether they should be conceived as nothing over and above patterns or collections of emotional episodes, or perhaps, as David Shoemaker claims about caring (I will argue, a species of sentiments), “a kind of conceptual-linguistic structural framework we overlay upon emotional reactions to render them intelligible” (2003, 94).

I argue in Chapter 1 that a reductive analysis of caring fails to capture our ordinary ways of thinking about sentiments. I choose to focus on caring because showing that it should be conceived as a sentiment construed as distinct from emotional episodes is already going some way towards showing that the neglect of sentiments is unwarranted. As the recent literature on caring (and love—I assume, a species of caring) testifies, we seem to take such a topic very seriously. So, if caring turns out to be a species of sentiment construed along the present lines, this gives us a strong reason to take sentiments seriously and thereby reject the common attitude of discounting sentiments as rather uninteresting or secondary. If, by contrast, all sentiments turned out to be at best interesting curiosities, entities that we can notice but leave aside at the profit of more important matters (I suspect the predominant trend among realists about sentiments), then this thesis would be at best a very local, very limited, contribution to the philosophy of mind.6 Needless to say, it purports to be more ambitious than that.

Sentiments should be conceived as distinct from emotional episodes. As I suggest in Chapter 1, however, the two kinds of entities seem to entertain a connection that we shouldn’t ignore; sentiments and emotions are not totally independent from each other. The trick is to find a category of entities that, while ontologically distinct from events, entertains a close relationship with them. As seen above in the context of regularities in nature, one possibility is to appeal to underlying dispositions or powers. The problem is that dispositional

---

6 One way sentiments could be relatively unimportant albeit interesting curiosities is by being mere neurological properties that are not a real part of our psychology. Innate temperaments may be an example. We’ll see in later chapters why the realist about sentiments might wish sentiments to count as genuinely psychological for her defence to be fully successful.
properties are often viewed with suspicion and themselves are claimed to be reducible to patterns or collections of events (some of which may be counterfactual). And it will be clear in Chapter 1 that the so-called ‘pattern’ theorist, while defining things like caring and love as dispositions to feel emotions, identify such dispositions with patterns of mental events. Agreeing with the claim that caring and love should be conceived as dispositions to feel emotions, but disagreeing, on conceptual grounds, that these should in turn be identified with patterns of emotions, I will as a result argue that an independently plausible non-reductive account of dispositions can help the realist about sentiments develop her view.

Chapter 2 will be dedicated to showing that dispositions in general should be conceived along realist lines. First, realism about dispositions captures our ordinary ways of thinking about them. Second, it provides more satisfying explanations than antirealism. In Chapter 3, I will provide a non-exhaustive list of the sorts of explanation that an appeal to distinct emotional dispositions (or sentiments) can allow, which will give the realist the upper hand.

In Chapter 4, I give an account of caring as disposition (or cluster of dispositions) realistically construed. On the account I propose, the dispositions that make up caring have properties—which I group under the general feature of ‘centrality’—in virtue of which caring is a special kind of sentiment. In order to assuage a worry that on my account sentiments end up being subpersonal causal forces, on a par with the fragility of the glass and the solubility of water, and thus not deserving of the label ‘psychological’, I then argue that there is no pre-theoretical reason to deny the status of intentional attitudes—of entities of a distinctively personal sort—to dispositional entities such as caring.

At the end of Chapter 4 and in Chapter 5, I argue that sentiments—in particular caring—should be part of an adequate account of character, and that recent scepticism about character traits is unwarranted because it is based on mistaken views about psychological dispositions.
Building on the point about the attitudinal aspect of (at least some) sentiments, I show in Chapter 6 that some sentiments are (or centrally involve) evaluative attitudes that can in principle be justified independently of the emotional episodes that they produce. Against the widespread claim that love is not a response to reasons, I show that paradigmatic forms of it can be conceived as at least partly cognitive states that are sensitive to evidence (albeit perhaps not as sensitive as many beliefs). I end the discussion by suggesting ways an account of emotion can allow sentiments to play a genuine justificatory role in the emotional episodes that they produce (contra a recent scepticism about the possibility\(^7\)).

Stimulated by a handful of pages from both Wollheim (1999) and Goldie (2011)\(^8\), this essay constitutes, to the best of my knowledge, the first sustained defence of the psychological category of sentiments.\(^9\)

\(^{7}\) Deonna & Teroni, 2012, 114.

\(^{8}\) Not to mention some enlightening discussions with Peter Goldie. Sigmund Freud may be another possibility, although I cannot claim to have been directly influenced by his writings, as my attention was explicitly brought to them only after having written most of this thesis. I do not wish to deny the possibility of an influence, perhaps via Wollheim, himself very much influenced by Freudian thinking. However, I would like to make it clear that a possible agreement with Freud on certain matters is compatible with a rejection of most of psychoanalysis' tenets.

\(^{9}\) Of course, ‘sentiment’ is often used to refer to the emotions themselves, so the reader should not take the term too seriously and rather take the aim of this thesis to defend the class of entities that ‘sentiment’ as I use the term purport to refer to. I hope though that calling them ‘sentiments’ is not entirely arbitrary, and that the term has useful connotations (such as the idea of depth) that the term ‘emotional disposition’ may not have.
Chapter One
Sentiments, Emotions, and Caring

1. Introduction

This thesis concerns the nature of sentiments, that is, dispositions of the mind to experience various emotional episodes in a variety of situations. Sentiments, sometimes also called ‘dispositional’ or ‘long-term’ emotions, include entities such as love, (long-term) jealousy, (long-term) anger, and irritability. When you love somebody, you are disposed to feel sad when that person is away, joyful when she is near you, relieved when you realise that she is safe, grief when she dies. When you love someone, you are disposed to feel all these things, making it sensible to define love in terms of such dispositions. Such dispositions, it seems, are necessary for there being an instance of love. If the relevant claims turned out to be false of you, if for instance you were not disposed to grieve when the person you love passed away, we could reasonably doubt that your love was genuine in the first place.

But what is it to be disposed to experience emotions in various situations? Is it simply to feel an emotion $e_1$ in situation $s_1$, $e_2$ in $s_2$, and so on? Is loving somebody merely a matter of feeling joy when she is around, sadness when she is gone, and so on and so forth? Or is it something more, or something different, or something distinct? Let’s call antirealism any view that takes sentiments to be nothing over and above a certain pattern of emotional episodes; all the explanatory work performed by sentiments is in reality work performed by emotional episodes or collections thereof. Two varieties of antirealism can be distinguished.

First, simple antirealism is the view, defended by Paul Ekman (1994) and (arguably) David Shoemaker (2003), that reduces sentiments to the mere
occurrence of certain emotions. On such a view, there are strictly speaking no long-term affective entities such as sentiments; talk of dispositions to feel emotions is instead shorthand for mere collections of mental events, namely the emotions themselves, collections that have no properties that would warrant either a non-reductive account or any room for sentiments in our theory of the emotions. If this is right, to love somebody is just to feel the relevant types of emotions in various situations.

An alternative account of sentiments does reduce them to collections of mental events—or more precisely to patterns of emotional episodes—and is thereby only committed to the existence to mental events and collections thereof, but takes such collections to play roles that are not played by individual emotions. Let’s call this view sophisticated antirealism. This view is antirealist for a similar reason the simple regularity theory of laws of nature that we saw in the introduction counts as antirealist; because, on this account, sentiments do not play a genuine role in explaining the occurrence of the corresponding emotional episodes, they are ultimately dispensable in our ontology of the mind.

With simple and sophisticated antirealism, contrast realism. This is the view that sentiments are not reducible to emotional episodes; the two sorts of entities are distinct. Furthermore, contrary to antirealism, realism takes sentiments to explain the occurrence of emotional episodes. Just as the realist about laws of nature, universals, or causal powers takes these entities to explain many if not all of the regularities instantiated between physical events, the realist about sentiments takes sentiments to explain many of the patterns of emotional episodes that people undergo over the course of their lives.\(^\text{10}\) One point of difference between the realist about sentiments and the realist about (say) laws of nature is that, while the realist about laws of nature may take laws of nature to be necessary entities, the realist about sentiments arguably takes sentiments to be contingent. Our psychology indeed could be in such a way that it didn’t include such entities. In addition, the realist need not be committed to the claim

\(^{10}\) As will become clear, I do not think the realist can reasonably claim that all emotional episodes are produced by sentiments. For instance, as we will see, it seems to be a metaphysical, if not empirical, possibility that emotional episodes can be directly induced into people’s heads.
that all sentiments are innate dispositions that may or may not be manifested in the various circumstances in which their possessor finds herself. My love for my friends is something I have acquired and not something that was there before I met them. Although it now explains the occurrence of many of the emotional episodes that I feel towards them, it didn’t explain the occurrence of the ones that I felt when I first met them. Plausibly, rather, the emotional episodes that I felt back then gave rise to my love, which is now a freestanding entity.

In this thesis, I will be defending realism about sentiments. I will argue that at least central cases of sentiments should be conceived as distinct from the emotional episodes to which they are related, in opposition with the reductive accounts introduced above. Central cases can be roughly defined as all the cases that we find very important to have and of which an account is particularly welcome. I take it that love is one such case, while irritability may not be. Furthermore, love may come in a variety of forms, including maternal attachment, friendship, and romantic love, and may itself fall under a more general kind of sentiment. This more general kind of sentiment, I suggest, is what we call caring.

While accounts of caring that do not specifically appeal to emotion can be found, it is quite natural to think that the two sorts of entities must be intimately related. If I care about my parents, I should be happy when they visit me; I should also miss them from time to time, worry if one of them were to get ill, and so on and so forth. If we were to take this observation seriously, we could justifiably define caring as a sentiment. In this chapter, I will argue that we should.

After a preliminary discussion on the nature of emotions and their relationship with cares, I will argue that a very plausible case can be made for the simple view according to which caring is a kind of sentiment. I will then introduce the two antirealist ways to interpret this claim and show that they are subject to counterexamples. I will end the chapter by suggesting that a realist
interpretation of the view is promising and by exposing the steps that will be undertaken to defend it.

2. Emotions and caring

2.1 Assumptions about emotions

After a long period of neglect, emotions are now a major topic in contemporary philosophy of mind. Naturally, several major theories of emotion are on the market. We have the early crude cognitivist account, according to which emotions are evaluative judgements or beliefs (e.g., Gordon, 1987, Solomon, 1993). Given the inadequacies of this account, less crude ‘neo-cognitivist’ accounts have been devised over the years. On such accounts, emotions are cognitive states that fall short of belief, including rather a construal of the world in evaluative terms (Roberts, 2003), or an evaluative thought about it (Greenspan, 1988). More recently, ‘perceptual’ theories of the emotions have been put forward as an alternative to cognitivist accounts, motivated either by the problems facing neo-cognitivist accounts or by similarities between emotions and ordinary perception. On such accounts, emotions should be conceived as direct (de Sousa, 1987, Tappolet, 2000) or indirect (Prinz, 2004) perceptions of value properties. Perceptual theories in turn face problems of

11 One problem with cognitivism is posed by the phenomenon of recalcitrant emotions, namely cases where one feels an emotion, such as fear, while making a conflicting evaluative judgement (“This dog is not dangerous”). According to cognitivism, this would be a case where one would be guilty of a radical form of irrationality (holding in mind two contradictory beliefs), a consequence some find unacceptable (e.g., Tappolet, 2000, Döring, 2009). Another problem concerns the subaltern status that the view gives to the phenomenology of emotions, construing it as a mere ‘add-on’ (Goldie, 2000, 40). See also Deigh (1994) for a classic critique of the cognitive theory of emotion.

12 Although she takes herself to be defending a non-cognitivist account of emotion, given that she defends emotions as primitive appraisals “in the sense that they occur without any conscious deliberation or awareness, and that they do not involve any complex information processing” (2005, 45), Jenefer Robinson would count as a neo-cognitivist on my understanding of the view.

13 See Brady (2009) for the claim that neo-cognitivist accounts fail to accommodate cases of recalcitrant emotions.
their own, at least taken as literally claiming that emotions are perceptions\textsuperscript{14} (e.g., Salmela, 2011, Whiting, 2012).\textsuperscript{15}

Given that I intend the argument of this thesis to be acceptable to most theorists of emotion, I will not commit myself to any particular theory of emotion. Instead, I will propose a list of features that I take to be necessary, though perhaps not jointly sufficient, for a mental state to be an emotion. Following are five plausible general claims about emotions that I will assume throughout this thesis. They provide a range of conditions that closely accord with both our intuitions about what emotions are and the way in which they are widely discussed in the literature. I remain however open to the possibility that there are other necessary conditions for a mental state to count as an emotion. I will also remain silent on the question of how the relevant features fit together.

The first general feature that is worth emphasising is that emotions are \textit{mental events}, more or less short-lived mental states that typically (if not always) involve phenomenology. The fact that, at least many times, emotions involve phenomenology is probably the reason why we tend to call them ‘feelings’.

Second, emotions are always directed at some things or other. When I am afraid, I am afraid \textit{of} something. When I am angry, I am angry \textit{with} someone. And when I am sad, I am sad \textit{about} something. To use a philosophical term of art, emotions are thought to involve \textit{intentionality}. Furthermore, they are often contrasted with \textit{moods}, such as a gloomy mood induced by sad music, which do not appear to be directed at anything, or at least not at things as specific as the objects of emotions.\textsuperscript{16} Thus emotions can be directed at anything that we can

\textsuperscript{14} As opposed to \textit{analogous} with perception. Roberts (2003), for instance, takes emotions to be analogous with perceptual states without claiming that emotions should be defined as perceptual states.

\textsuperscript{15} One problem with perceptual accounts is that they fail to accommodate the thought that emotions are things for which we can provide reasons, given that perceptual states do not seem to be the kind of thing to be justified or unjustified (Deonna & Teroni, 2012, 69; see also Brady, 2010).

\textsuperscript{16} Some philosophers (Goldie, 2000) take moods to be directed \textit{at the whole world}. This view, however, seems to some (e.g., Deonna & Teroni, 2012, 13) to have the implication that all instances of moods are inappropriate, incorrect, or ‘unfitting’ (the alternative view being that
think of: people (my anger towards Mary for having hurt my feelings), animals (my fear towards a dog with big teeth), artworks (my admiration towards a Monet painting), events, states of affairs, and so on. An advantage of the notion of intentionality is that it also captures the fact that emotions can be about things that do not exist, haven’t occurred yet, or are no longer existing, such as the sadness I feel for the supposed death of someone who, unbeknownst to me, hasn’t really died, or the nostalgia I feel towards things that are no longer of this world.

Third, emotions are *cognitive* in the sense that they present the world as being a certain way. For instance, sadness, fear, and anger present the objects to which they are directed as involving some loss, some danger, and some offense for the subject undergoing it, respectively. For each emotion type, a general value property (loss, danger, offense, etc.) can be identified as the kind of property that is attributed by it; in other words, emotion types can be individuated by means of the value properties that they attribute to their objects. Furthermore, such properties, sometimes called *formal objects* (Kenny, 1963), provide accuracy conditions for each emotion type. Sally’s fear at a dog is accurate just in case the dog in question constitutes a threat to her well-being, inaccurate otherwise. We sometimes speak of emotions being appropriate or inappropriate, as when we say that it is inappropriate to feel rage at a minor offense. However, given that the notion of appropriateness is sometimes understood in terms other than accuracy, I will avoid appealing to it in this chapter.17 Following D’Arms and Jacobson (2000), I will reserve the predicates ‘fitting’ for any emotional episode that is directed at an object instantiating the episode’s formal object and ‘unfitting’ for any emotional episode that is directed at an object that fails to instantiate the episode’s formal object.

---

17 Greenspan (1988), for instance, construes appropriateness as a measure of how justified an emotion is. One can also think of moral and prudential ways to construe the notion, as when we say that it is (morally) inappropriate to laugh at an offensive, yet funny, remark or that it is (prudentially) inappropriate to feel fear at an animal if the animal approaching us can sense it. See D’Arms & Jacobson (2000) for a discussion.
Emotions are not simply things that can be fitting or unfitting. Emotions—and this is the fourth feature—are subject to epistemological standards. Reasons can indeed be given for the evaluation that an emotion makes. Why should I feel angry at Frank’s remark? Why should I take it as offensive? For one thing, I may have heard it. If my anger were instead based on the testimony of an unreliable friend, it would be far less justified. I may also be justified in feeling angry because I have evidence, such as the tone of his voice, that Frank intended his remark to hurt me. The content of the remark may have conventional connotations that make the remark an offensive one as well. The list of pieces of evidence can go on. In that respect, emotions differ from perceptions in that the latter are typically not thought to be supported by evidence (Deonna & Teroni, 2012, 7). Moreover, like belief states, emotions can be accurate (or ‘fitting’) even if they are not justified (or based on little evidence)—I may have no good reason to be angry with Frank even though he said some nasty things about me behind my back (which would make an angry response fitting). And they can be perfectly justified even if they are inaccurate (or ‘unfitting’)—although I would be justified in feeling fear at the view of what looks like a very threatening animal, my emotion would be unfitting if the animal turned out to be a very lifelike statue.

The fifth and last feature of emotions that I would like to emphasise is their connection with behaviour. First, emotions are states such that there is something it typically looks like to be in them. Fear, anger, sadness, and shame can all be inferred from certain facial expressions and bodily postures. Although one may attempt to conceal such expressions, it is undeniable that central cases of emotion involve a set of characteristic behavioural tendencies. Furthermore, emotions not only can trigger such involuntary behaviours, but also produce desires for action and ultimately the actions themselves. In short, emotions can both motivate and explain action (e.g., Tappolet, 2010). One’s anger at someone else can motivate a violent act, and one’s fear at a dangerous animal can motivate the desire to flee. Finally, it is plausible to say that emotions’ capacity to recruit and direct one’s attention in a way that makes features of the
environment perceptually salient contributes significantly to emotions’ motivating power (e.g., de Sousa, 1987).

2.2 The connection with cares

Now that we have a firm, albeit admittedly incomplete, grasp on the notion of emotion, we can motivate the thought that emotions are intimately related to *caring*.

First off, there is some evidence suggesting that the two notions are conceptually related in some way. If you say that you care about X, or verbally judge that X is good or important in a way that suggests that you care about it, but you never feel anything about X (or things suitably related to it), then we have reasons to doubt your sincerity. As David Shoemaker says, “if one actually cares about Y, one is expected to have emotional reaction X, depending on what happens to Y. If one does not, then it seems that one’s claims of caring are disingenuous.” (Shoemaker, 2003, 93) And this seems to accord with people’s judgements. In *The Emotional Construction of Morals* (2007), Jesse Prinz mentions an unpublished study of his where subjects were asked to consider “two scenarios in which emotions are pitted against verbal behavior.” (99) Here’s the summary of the experiment:

In one scenario, one student in a fraternity insists that there is nothing morally wrong with smoking marijuana, but he feels disgusted when he sees his fraternity brothers smoke and ashamed when he himself smokes. In the other scenario, a student in a fraternity tells his brothers that smoking marijuana is morally wrong, but he never gets disgusted at them when they smoke, and he would not feel ashamed if he smoked. In each case, subjects are asked whether the

---

18 I construe conceptual relations broadly as including typical associations between concepts. So, if it turns out that our thinking about caring is regularly, but not always, accompanied by thinking about the emotions, and that the two are not related by something as strong as conceptual entailment, this will nonetheless count as a conceptual connection in my sense.
student’s moral values are reflected by what he says or what he feels. Subjects in both conditions were much more likely to say that the value corresponded to the emotions; the student who feels ashamed of smoking marijuana is morally opposed to it, and the shameless student is not. This suggests that folk morality draws a conceptual link between values and emotions. (Ibid.)

Another piece of evidence for a connection between emotions and cares comes from our practices of making sense of other people’s emotional episodes. When someone else is reacting emotionally to a particular situation, we can ask for an explanation of its occurrence. We may of course point to relevant features of the situation, such as the fact that a remark was offensive. But when we, as outside observers, do not find in the situation features that would make the emotion fitting or justified, we can search for features of the subject that would explain her reaction. The need for this sort of explanation is particularly acute when we have no idea why anyone would react in this way in the situation in question. Consider, for instance, the case of an old person who gets angry every time she is offered help. We may wonder why she gets so angry when being offered help is generally a good thing. There are many reasons she could get angry. For instance, it could be a mere quirky fact about her that she gets angry in just these circumstances, on a par with phobic responses perhaps. Another possible explanation, however, is that she cares about being fully autonomous and as a result finds it offensive (by feeling anger, given that to be angry is, at least, to evaluate something as offensive) when someone else offers their help. Knowing that autonomy is important for her, as a result, we can make sense of the pattern of emotional responses that she instantiates in the relevant situations.

Additionally, people’s emotional responses are often used as evidence for the presence of a care. In our example, the fact that the person reacts in the systematic way she does is evidence that she cares about something that is related to the object of her emotion. In particular, regularities or patterns in emotional responses can be made sense of by an appeal to the fact that the
subject cares about certain things and not others. In order to know what you care about, I need to know how you would react to various events, people, and situations over time. For instance, one’s joy upon meeting an otherwise dangerous dog could be explained by one’s love for dogs or by one’s caring for dangerous activities; and in order to know which one is the case, we need to see what activities they engage in over time and their emotional reactions to them. This may suggest that we take the presence of a certain pattern of emotions as necessary for the presence of a care. In line with this point, Shoemaker writes:

Genuine caring about something involves a package deal: one must, along with the possibility of joy (and other positive emotions), accept the possibility of distress (and other negative emotions) when things are not going well with the cared-for object in order for one truly to be said to care for it in the first place. To claim to love something only when things are going well for it seems to be no love at all. (92)

One last piece of evidence for a close link between emotions and cares is the fact that a change in what one cares about seems to imply a change in how one would react emotionally to the world. The fact that Betsy does not care about the welfare of animals anymore, and the fact that she now cares about French cuisine, explain both why she is no longer disgusted by meat and why she instead feels joy at the view of a piece of filet mignon.

There is a connection between emotions and caring that is hard to deny. In the following section, I will put forward a general account of their relationship that purports to do justice to all of the facts that we have just uncovered.
3. The Simple View of cares

Before proposing an analysis of caring, I’d like to introduce some terminology. Following Helm (2001), I’ll call the object of a care its focus and the particular things towards which the relevant emotions are directed and that are suitably related to the focus the emotions’ targets (which can sometimes coincide with the focus). For instance, the anger I feel towards you after having dirtied my favourite shirt has you as its target and my favourite shirt as its focus (or my favourite shirt is the focus of the care my anger is an expression of). One word on the phrase ‘suitably related’ is in order. I admit that I have not found a way to cash it out in a precise manner, but a not entirely unsatisfying way to elucidate it is by saying that an emotion’s target is suitably related to a focus just in case the subject is having the emotion towards for the sake of. I will take the locution ‘for the sake of’ as primitive, as I am not aware of any way to define it in a non-circular manner. One thing to notice though is that the relation that obtains between a target and a focus depends on a psychological factor—the subject’s responding to one for the sake of another—and not merely on objective factors such as an appropriate causal relation between the two entities (which, for one thing, can sometimes be identical). I doubt anyway that any general objective factors, as opposed to idiosyncratic ones, could be appealed to here, as it is clear that virtually anything can be a target towards which we experience an emotion for the sake of something else that we care about. For instance, I can feel joy thinking about the number 459 for the sake of a departed mathematician friend because, for some reason, it used to be his favourite number.

We can now propose the following (partial) analysis of caring:

---

19 Strictly speaking, on Helm’s account, the object of a care is the focus of the various emotions that constitute it (as we will see). There is no harm however in calling the object of a care its ‘focus’, especially if we do not wish to identify caring with a pattern or complex of emotional episodes. In the following, I’ll speak both of the focus of emotions and of the focus of cares. The focus of an emotion, on my account, is just the focus of the care it is expressive of (if it is expressive of a care at all), where the expression ‘being expressive of’ is for now a placeholder for whatever relation turns out to obtain between cares and individual emotions.
Simple View of Cares: X cares for focus f only if X is disposed to feel a range of emotions for the sake of f over time and in particular circumstances. In other words, caring is a kind of sentiment.

Before discussing the various ways we can interpret the Simple View, we can note that it has two general advantages. On the one hand, it accounts for the various links between emotions and cares that we have identified in the previous section at least as well as the major alternative accounts. On the other hand, it is not subject to problems that the major alternative accounts face. It should be noted however that the reasons I will give to opt for the Simple View and reject alternative accounts are not meant to be conclusive ones. There are many accounts of caring on the market and it would take us too far afield to assess each of them individually. In the following, therefore, I will mainly provide a positive case for the Simple View, leaving a full defence of it against alternative accounts for another occasion.

Two alternative general accounts of caring can be distinguished. The first takes the tokening of an evaluative judgement to be necessary for caring (Watson, 1975, Buss, 1994, 1997). On such an account—call it the judgement account—I care about X only if I judge that X is valuable, where to judge that p is to take p to be true (to ‘assent’ to it). The second kind of account—call it the higher-order account—takes our ability to attend to and identify with our own mental states to be necessary for caring (Frankfurt, 1971, 1999, Velleman, 2000, Bratman, 2007). For instance, according to Harry Frankfurt, one cares about something just in case one desires to desire that thing. More precisely, on his account, if I care about X, I will desire that certain first-order desires of mine be my will, that is, that they be effective in motivating me to act in relevant ways.²⁰

---

²⁰ It may be wondered why Frankfurt does not define cares in terms of simple, first-order desires. The answer is that if to care about X is simply to desire X, then conflicts between cares and desires would turn out to be mere cases of conflicts between first-order desires. For instance, a drug addict may strongly desire to take a fix while caring about stopping. In such a case, it seems intuitively wrong to say that the taking of a fix and the quitting of drugs are equally important states of affairs to him.
How does the Simple View fare against these accounts? The first thing to be said in favour of the Simple View is that it is at least as much in sync with the thoughts introduced in the previous section as the judgement and higher-order accounts, giving us prima facie reasons to take it seriously. First, let’s compare the Simple View with the judgement account with respect to these thoughts. It is important to point out there is no reason to deny that the judgement account can accommodate the idea that cares can explain the occurrence of emotions (by saying that evaluative judgements constitutive of caring produce the relevant emotions) and that changes in emotional responses correlate with changes in caring (by saying, as Smith (1994) for instance does, that a change in evaluative judgements tend to produce changes in motivation, broadly construed so as to include emotions). Naturally, the judgement account could also accommodate the thought that emotions can provide evidence for the presence of a care: returning to our example, the fact that the old person feels angry when offered help can be explained, on the judgement account, by the fact that she judges that autonomy (say) is a good thing. Where the judgement account seems less plausible than the Simple View is in accounting for the thought that emotions and cares are conceptually related in a way that judgements and emotions do not seem to be.\(^{21}\) Given that the conceptual link between emotions and caring seems more robust than the conceptual link between emotions and evaluative judgement, we therefore have a reason not to identify cares with evaluative judgements and to instead identify cares with entities more closely related to emotions. The judgement theorist could however object that the fact that someone is not moved after having made his judgement is evidence that the judgement was not sincere in the first place, or she could explain the case in question as a case where certain obscuring factors (such as peer pressure) are at play in the situation (Smith, 1994). A proponent of ‘motivational internalism’ about evaluative judgement—according to which evaluative judgements conceptually entail motivation—does seem to have the resources to explain the conceptual link between caring and emotion if she were to accept the view that caring is evaluative judgement. I do not have any objection to this line of

\(^{21}\) Recall Prinz’s case of the student who, while judging that smoking marijuana is wrong, is never disgusted by it.
response, other than that motivational internalism may be more controversial, and therefore in stronger need of argument, than the claim that caring and emotion are conceptually related, giving some prima facie evidence against the identification of caring with evaluative judgement. (It just seems true to me that one can judge that something is good while not caring at all about it, but this may be only me.) At any rate, I find myself in the dialectically comfortable position of holding a view whose success is not hostage to the fortune of a thorny metaethical debate. The Simple View is, as a result, only somewhat more attractive than the judgement account in accounting for the thoughts introduced in the last section. However, we will shortly see that there are further reasons to prefer the Simple View.

Higher-order accounts may also have some resources to explain the thoughts introduced in the previous section. Let’s consider Frankfurt’s account in terms of second-order desires. Desires, we all know, are things that can be satisfied and frustrated. When they are satisfied, we tend to experience positive emotions, and when they are frustrated, we tend to experience negative emotions. Thus, emotions can be used as evidence for the presence of desires. A change in desires, moreover, tends to produce a change in emotional reactions. I used to desire to be a professional skateboarder, and used to feel frustrated when the prospects of becoming one were diminished (e.g., when I had no time to train); but given that I don’t have that desire anymore, the fact that the prospects of becoming one are dim now does not upset me. So it seems that the Simple View and at least one higher-order account can both do justice to some of our pretheoretical thoughts about the relationship between emotions and caring. The choice between them will have to be made on independent grounds. Note, however, that Frankfurt’s account need not be incompatible with the Simple View. All the Simple View says is that caring is an emotional disposition of some sort; it doesn’t say what sort of disposition it is. If desires, including second-order desires, are to be understood as sentiments, therefore, they could in principle constitute cares. As we will see in later chapters, this is a possibility
that should be taken seriously.\textsuperscript{22} For now though, I am assuming that the Simple View is in competition with all higher-order accounts, including Frankfurt’s.

It looks like all the theories reviewed so far—the Simple View, the judgement account, and the higher-order account—are roughly equally plausible. I will now introduce two considerations that should tip the balance in favour of the Simple View. The first consideration should cast doubt on the judgement account only. The judgement account takes evaluative judgement to be necessary for caring. It thereby explains the undeniable link between caring and emotion in terms of a supposed link between evaluative judgement and motivation. In order to test the plausibility of this hypothesis, let’s consider a case where a subject judges that something is (say) bad yet feels emotions characteristic of caring towards it (and things related to it). An abused wife, for instance, can judge that her husband is unworthy of her love yet care deeply about him. In such a case, we intuitively want to say that, although she judges—perhaps even knows—that her husband is bad for her, the subject can still care about him. The question now is, is this case akin to a case of conflict in judgements or is it akin to a case of a conflict between a judgement and something else (as in cases of recalcitrant emotions)? My intuition here is that the case does not—or at least need not—involve a conflict in judgements. It is indeed more akin to a conflict between a judgement and a phobic response than it is to a conflict in judgements.\textsuperscript{23} There does not seem to be any good reason to deny that the abused wife could make the all-things-considered judgement that she should leave her husband and yet still care about him and feel emotions that only someone who cares about him would feel. Caring, it seems, is more of a matter of emotion than the judgement theorist makes it out.

\textsuperscript{22} Given the problem that will be shortly raised against higher-order accounts, the possibility that a desire-based account of caring should be taken seriously is one that does not take higher-order desires to be necessary for caring. First-order desires, at least those that count as ‘central’ (see Chapter 4), may be sufficient.

\textsuperscript{23} Another phenomenon to which conflicts between judgements and cares may be akin to is the case of gender or racial bias where the subject may judge that (say) equality between genders should be promoted yet tends to react in a way that contradicts his judgement (e.g., by not taking seriously enough his female colleagues).
The second consideration that I think points in favour of the Simple View is that both the judgement account and the higher-order account place severe constraints on the kind of subject that is capable of caring. In a recent paper, Agnieszka Jaworska (2007) convincingly argues that individuals, such as human infants and Alzheimer’s patients, can have the ability to care while lacking some of the cognitive functions necessary for the formation of judgements and second-order mental states. The evidence she cites is both anecdotal and scientific. On the anecdotal side, she mentions cases where “a typical two-year-old (...) insists on doing many routine tasks—from eating to tying her shoe-laced—by herself.” (530) She gives the example of the son of a colleague of hers who “took very seriously the task of blowing out all the candles on his third birthday; he prepared for the event, got very frustrated with his initial not-so-successful attempts, triumphantly announced his eventual success, and reminded his patients about the feat for days afterwards.” (531) On the scientific side, she mentions “reports of patients in moderate to severe stages of [Alzheimer’s] whose complex patterns of behavior strongly suggest that they care about helping others, about participating in meaningful activities, about contributing to science, about maintaining their self-image and reputation, and the like.” (ibid.) As a result, given that both young children and certain clinical patients presumably lack the cognitive capacities necessary for judgement and second-order mental states, but that they possess the capacity for caring, we shouldn’t think that caring requires those mental states.

Notice that even if it turned out that the human infants and Alzheimer’s patients who have the capacity to care did possess the relevant cognitive capacities, Jaworska’s argument would still have some force. For if, by hypothesis, we are confronted with an individual lacking the relevant capacities yet would still be willing to attribute to them caring attitudes, this means that caring attitudes must be distinct from the relevant capacities. For all I know, even certain non-human animals genuinely care for their young, and that even if they may possess rational capacities that are too minimal to produce judgement and higher-order mental states of the sort required by the judgement and higher-order accounts. If
a view of caring does not rule out *a priori* such cases, then this is the one that we should choose, especially if all other things seem to be equal.

We can conclude that the Simple View should be taken as a serious alternative to the two popular ‘intellectualist’ accounts that we have discussed. I do not wish to pretend that the reasons given in favour of the view are conclusive ones, but they are sufficient to motivate its development. In the following section, I will introduce two ways to understand the Simple View that can be drawn from the philosophical literature on caring. Both views, I will argue in Section 5, should be rejected.

4. Antirealist interpretations of the Simple View

Now that we have seen some immediate advantages of the Simple View, we must spend some time trying to understand it in more details. If to care is to be disposed to feel a range of emotions in various situations, what is it to be ‘disposed to feel a range of emotions in various situations’ (or ‘to be disposed to feel’ for short)? This question is rarely asked in the literature on emotion, as it may seem that the answer is obvious. Two similar but different interpretations of the Simple View can be found in the literature, which I will introduce in this section.

Both interpretations of the Simple View to be considered analyse cares (and more generally sentiments) in terms of patterns (series, sequences) of emotional episodes, but differ in their take on the question whether such patterns are both necessary and sufficient for caring. For now, let’s try to formulate the view that defines cares in terms of patterns in a way that is neutral between the two interpretations. On a first try, we can say that to be disposed to feel is to instantiate a tendency to feel certain emotions over time, where to instantiate a tendency is to instantiate a certain regularity in feeling. To be disposed to feel for X, as a result, is to feel a certain kind of emotion in C1, another kind in C2, and so on, where C1 and C2 are situations that are occurring or have occurred in the past. Call ‘pattern-based’ any account that takes the instantiation of a pattern
to be necessary for caring. The view under consideration is pattern-based, but understands the instantiation of patterns to involve actual emotions only:

*Actualist pattern-based account of cares:* A cares for B only if A is disposed to feel for B, and to be disposed to feel for B is to have exhibited a range of emotions over time.

Although the view may sound natural, it doesn’t seem to capture a crucial feature of cares, namely that the having of them makes certain emotional responses *possible.* It may be true that A cares for B, where caring for B implies that A would grieve if B were to die, but it is not because A may never be able to grieve for B (because A may well die before B) that A can be said not to care for B after all. The actual instantiation of the entire range of emotions characteristic of caring is not necessary for caring\(^{24}\); what matters is that the subject *would* feel in certain ways:

*Conditional pattern-based account of cares:* A cares for B only if A is disposed to feel for B, where to be disposed to feel for B is to exhibit a range of emotions in certain circumstances (some of which may never present themselves).

Given the implausibility of the actualist version of the pattern-based account, when I speak about the ‘pattern-based’ account in the following, I only refer to the conditional version.\(^{25}\)

Another way to put the conditional version of the Simple View is to say that A cares about B just in case A is disposed to feel for B, where A is disposed to feel for B just in case, were A to find himself or herself in a range of circumstances C, A would undergo a range of emotions whose focus is B. One thing to notice

---

\(^{24}\) Though the instantiation of *some* emotions presumably is necessary, especially for the pattern-based account that takes caring to be a conceptual device used to make sense of other people’s emotions over time. See below.

\(^{25}\) Another way to formulate the pattern-based account is to say that a care is an ‘emotion complex’ constituted of emotional episodes and (possibly) other affective phenomena, unfolding over time (Jaworska, 2007). I’ll take the two views as equivalent, even though the emotion complex view may suggest an actualist reading.
is that it shouldn’t be just any emotions that one feels in a range of situations in order for one to care for a given thing; in order for one to be said to care about a particular thing—in order for a collection of emotions to constitute a pattern of the relevant sort at all—one must have emotions whose focus is that thing. If one picked at random a thing in the world and wondered whether I care for it, one would therefore have to check not only whether I would feel certain emotions in various situations involving the object in some way, but also whether the emotions in question have as a focus the object itself—whether the targets of the emotions are suitably related to the object. If the focus of the emotions is not the object, then I cannot be said to care for it; at best, I care about something else in the vicinity, but not that object. Moreover, in order for an object to constitute the focus of a care, how it fares (positively or negatively) must correlate with the type of emotion one feels (positive or negative). For instance, if I care about a friend of mine, I should feel joy when his life is going well and sadness when it is not. So, patterns of emotional episodes constitute a care in virtue of the episodes’ having what I called a common focus.

Now, let’s look at the two versions of the pattern-based accounts. Both take a pattern of emotions of the relevant sort to be necessary for caring. Both also construe sentiments as ontologically dependent on such patterns. However, only one of them takes the relevant pattern to be also sufficient. On this version, caring is a mere pattern of emotions, in the sense that it is a pattern of the relevant sort and nothing else. In other words, no other conditions are necessary for caring, on such a view. This is what we can call the ‘simple’ antirealist account of caring. Recall Ekman’s remark, mentioned in the introduction of this thesis, that talk of ‘long-term’ emotions is merely shorthand for series of brief emotional episodes (Ekman, 1994). To him, when we claim that Anna loves her father, we are in effect saying that she is disposed to feel a range of emotions towards her father (and related things), understood in terms of the pattern-based account above. David Shoemaker seems to agree when he claims that

caring is a kind of conceptual-linguistic structural framework
we overlay upon emotional reactions to render them
intelligible…. Talk of caring is simply a way of referring to the range of emotional reactions one is expected to have with respect to the fortunes of the cared-for object. To care is thus to have the disposition to feel… (2003, 94)

On such a view, therefore, the notion of caring turns out to be a conceptual device we use in order to make our emotional responses intelligible to ourselves, to approach them in a way that has some cognitive significance.

The simple antirealist therefore takes cares to be mere patterns of emotional episodes. The sophisticated antirealist disagrees, and argues that further conditions must be met in order for a pattern of the relevant sort to constitute an instance of caring. According to Bennett Helm (2001), patterns of emotional episodes must involve members that are rationally connected. When one feels a given emotion, Helm claims, one commits oneself to two sorts of commitment. First, we have ‘transitional commitments’, whereby the having of an emotion of a given type at a time commits oneself to the having of an emotion of a different type at a later time. For instance, my hoping that it will not rain tomorrow commits me to feeling relief if it finally doesn’t rain. Second, we have ‘tonal commitments’, whereby the having of a positive emotion in a certain situation commits one to feeling the corresponding negative emotion in relevant counterfactual situations. For instance, if one feels joy because it doesn’t rain, one is thereby committed to feeling sadness or disappointment in a counterfactual situation where it does.

What distinguishes the simple antirealist and the sophisticated antirealist, therefore, is the fact that the former takes ‘mere’ patterns to be necessary and sufficient for caring while the latter only takes such patterns to be necessary,

---

27 As pointed out by Michael Scott (personal communication), this requirement looks odd. How could I be committed to having a feeling in a situation that did not arise? My assessment of the view, however, will not depend on the particular details of it, so I will leave this worry aside in the following.
claiming that the relevant patterns must satisfy further conditions. Notice though that the sophisticated antirealist account includes the simple antirealist claims plus further claims. All the cases the sophisticated antirealist counts as cases of caring are cases that the simple antirealist will count as cases of caring as well, but not vice versa. In other words, the set of cases that the sophisticated antirealist accounts for is a subset of the set of cases that the simple antirealist accounts for. As a result, counterexamples to the former should constitute counterexamples to the latter as well. In the next section, I will put forward cases that the sophisticated antirealist—hence the simple antirealist—should have a hard time accommodating, giving us reasons to look for an alternative way to develop the Simple View.

5. Counterexamples to antirealism

In the following, I will attempt to show that it is intuitive to think that caring is in fact something over and above patterns of emotional responses to which it seems related. First, I’ll show that there are cases where a pattern is present but where we are reluctant to attribute a genuine care. Second, I’ll show that there are cases where a care is plausibly present but where there is no corresponding pattern. The two kinds of cases jointly constitute an argument to the effect that the instantiation of a pattern is neither necessary nor sufficient for caring.

5. 1 Patterns without cares

Before introducing a few counterexamples, I would like to make a methodological point. As I said, the pattern-based account, by identifying cares with patterns, makes a conceptual claim about cares. Construed as a claim about concepts, the claim as a result can be put to test by means of various imaginary cases. The account after all is not making an empirical claim about what

---

28 Another way to look at the distinction between simple and sophisticated antirealism here is by saying that they differ in their account of the kind of pattern that is necessary and sufficient for caring. Simple antirealism construes it as a mere collection of emotional episodes with a certain focus, while sophisticated antirealism construes it as a collection of emotional episodes with a certain focus and subject to rational norms. I take it that nothing substantive hangs on whether we look at the distinction in this way instead of the one described in the main text.
properties cares happen to instantiate in the actual world; it makes a stronger claim about what it is to care about something. All we need to determine, therefore, is whether there are both cases with patterns without cares and cases with cares without patterns; in short, cases where we find a dissociation between patterns and cares. In this section and the next, I will show that there are such cases and that we thus have reasons to deny the two directions of the bi-conditional ‘X cares for Y if and only if X instantiates a pattern of emotions P (with rational interconnections)’.

Consider the following scenario, given by David Pugmire:

Suppose there are two people of contrary character. There is Nick, who is ill at ease with others and over-anxious to please. He can’t abide silence and has to maintain a barrage of constant, random banter. Rick, on the other hand, is ruminative, laconic and doesn’t suffer fools gladly. When their professional circumstances force them together Nick finds Rick perversely unforthcoming and even hostile—thankless company and, frankly, intimidating. For his part, Rick finds Nick an irritating twit. Not surprisingly, they bring out the worst in each other: insistent nervous chatter confronts bristling silence. Their working relationship suffers. However, their Company has a resourceful Personnel Department. Without their knowledge the Department medicates both men with a new psychotropic drug called Amity. Their perceptions of one another remain, but their mutual responses are wonderfully transformed. Thus, despite Nick’s admittedly pathetic foibles, Rick warms to him. Irritation and disdain yield to bemused indulgence. On the other hand, Rick’s ominous silences no longer prevent Nick from dubbing him somehow likable. As if by grace, an unaccountable fondness arises on both sides, and the working relationship prospers. (Pugmire, 1994, 106)
The question that we should ask ourselves now is, do Nick and Rick really care for each other? In the scenario, we have, it seems, the relevant kind of pattern, both with a common focus (the other) and plausibly satisfying the requirements exposed above (transitional and tonal commitments). Yet, given that it is the pill that is doing all the relevant work, and not, say, a deeper part of the subjects’ psychology, we can confidently say that Nick and Rick do not really care for each other.

The pattern theorist may object that, given that it is unclear what work the pill is doing here, it is unclear whether we should be confident in our judgement. The pill, after all, must be able to track certain relevant facts in order to be activated at all. In order to avoid this sort of mystery, let’s imagine that instead of the pill, we have a high-tech device that is implanted into the characters’ heads and that, whenever they are in the relevant cognitive states, triggers the relevant emotions. For instance, whenever it detects a perceptual belief that the other is in the vicinity and that he is in a state of distress, it triggers friendly compassion directed at him. And we can assume that the device is so sophisticated that it triggers not only the right sort of emotions, but does it in a way that feels natural to the protagonists. My take on this case is that there is at the very least ground for saying that it is an open question whether Nick and Rick genuinely care for each other, suggesting in turn that the instantiation of a pattern is not sufficient for caring.

A variation on the case above is where it is not a pill or electronic chip that is responsible for the ‘inauthentic’ emotions, but the subjects themselves. Cases of ‘sentimentality’ are cases where an emotion is partly motivated by the desire to have it (e.g., Tanner, 1976). I can for instance ‘work my self up’ in order to be angry at what I take to be a harmless remark if I believe that everyone else thinks that it is an insult and I do not want them to think I’m a coward (Dilman, 1989). Other cases include working one’s self up to grieving for someone’s death for which we have no reason to grieve (as opposed to merely find it sad and unfortunate), as in the death of certain celebrities (Milligan, 2008).
The pattern theorist could reply at this point that such emotions do not constitute cares because they are not rationally interconnected. One’s grief for a celebrity is indeed such an isolated case that one cannot be said to really care for her. My reply is that I don’t see why one cannot (at least in principle) work one’s self up to an entire pattern of emotional episodes that happen to satisfy the pattern theorist’s requirements. Imagine that John would like to love Kate (because he married her for financial interests, say), but finding himself not to feel anything for her, he decides to attempt to feel all the emotions he would have if he loved her. Of course, at first, he would have to associate her with other things that he loves, such that thinking about such things may trigger the relevant emotions. Still, it is possible that he manage to have the relevant emotions directed at her. By managing to feel the relevant emotions, does John ipso facto manage to love Kate, or is it a further question whether he does? Although the fact that John feels in certain ways is clear evidence that he loves Kate, it doesn’t seem impossible that he is completely self-deceived and that, were he not keeping working himself up, he would not feel the relevant emotions anymore. The pattern theorist, however, is forced to say that, if a pattern of the relevant kind is present, a care is thereby present.

One strategy that the pattern theorist could use is to argue that in all the cases that I have presented, the relevant affective states are not genuine emotions. If it is not genuine grief that I feel for the death of a celebrity that I barely know, the pattern theorist must say why, and more importantly explain what is missing for a genuine care to be there. Of course, there are differences between the self-induced states and the more ‘natural’ ones, but these differences must include differences in facts about caring in order for the pattern theorist’s claim to be plausible. She must indeed say what is missing in either the content of the emotions or in their connection with other mental states (i.e., the two criteria for collections of emotions to constitute a pattern and therefore a care). Alternatively, the pattern theorist could add further conditions in order to distinguish between patterns that are cares and patterns that are not cares.
A similar point can be found in Seidman (2009). According to him, the pattern theorist fails to accommodate a particular class of cases where a pattern is present but where no care can be attributed, namely cases of patterns of ‘superficial’ emotions. For instance, being irritable after having drunk too much coffee, I can be angry with you for humming (an activity which, let’s suppose, I normally do not find annoying), relieved when you stop, really annoyed when you resume it, and so on. In such a case, there is pattern of emotions with a common focus (you or the humming) and rationally interconnected (from anger to relief), but, it seems, no care.

The pattern theorist can reply to this worry by suggesting that a pattern constitutes a care only if it includes certain types of emotion, ruling out cases of ‘superficial’ patterns such as the one above. The requirement could be that patterns of emotions higher in cognitive sophistication than reflex-like emotions such as anger, surprise (or ‘startle’) and fear (what Damasio calls ‘secondary emotions’, 1994, 136) are necessary for caring (Jaworska, 2007, 557, Shoemaker, 2003, 93-94, ft. 13). Secondary emotions include shame, guilt, hope, grief and regret. According to the proponents of this claim, although certain emotions can be construed as forces that can in principle be induced by artificial factors (such as a pill), this is not true of the class of secondary emotions. “We can perfectly well imagine being overcome by any particular primary emotion such as anger, or fear, or disgust, so that one considers the emotion external to oneself. But it is hard to imagine one’s grief or gratitude as external forces that simply occur within oneself.” (Jaworska, 2007, 557) Hard to imagine, perhaps, but is it really an impossible state of affairs? Is it really impossible that the relevant complex emotions cannot be induced by artificial means?

Recall that the pattern theorist makes a conceptual claim about cares, not merely an empirical one. Of course, it may be the case that to each pattern of secondary emotions we happen to instantiate, there is a relevant care. However, if we are

---

29 Bennett Helm (2001, 103) makes a similar distinction between reflexive emotions (shame, guilt, etc.) and non-reflexive emotions (anger, surprise, etc.). In this chapter, I take the two distinctions to be roughly the same.
to accept the possibility of the cases above (the Nick and Rick case and the cases of sentimentality), we shouldn’t rely on our intuitions about the possibility of dissociating ourselves from secondary emotions in order to conclude that cares must be identified with patterns. After all, it seems possible to feel emotions that we identify with (by taking them as fully our own) but that are nonetheless external to us in the relevant sense. We would just be wrong in identifying ourselves in this way; and this may actually be partly what is tragic about such cases.

Two remarks are in order before moving on. The first thing to say is that I do not wish to suggest that the pattern theorist is without resources to accommodate the counterexamples that I have given. There could be details, more or less subtle, that have been left out in the description of the scenarios, and that could show that the pattern-based account is worth preserving. There could also be hidden controversial assumptions. One such assumption may be that an emotion can in principle exist—and thereby be induced—on its own, without being related to other mental states. As a result, I do not wish to put too much weight on the counterexamples to the pattern-based theory that I’ve given so far (as opposed to the one to come). They should be seen as at best requiring the pattern theorist to give us her version of the story regarding the relevant cases, a story that may or may not turn out to be acceptable. The second thing I would like to say is that there is a sense in which the people in the scenarios who instantiate ‘alien’ patterns of emotions may nonetheless be said to ‘care’ in some way for the apparent focus of these emotions. The point however is that this is not, it seems, what we typically mean when we speak about people caring about things in a deep sense. At best, therefore, the pattern theorist can account for caring in an *attenuated* sense (a similar way the empiricist about laws of nature is free to call regularities in the physical world ‘laws of nature’).³⁰

---

³⁰ One could for instance distinguish between deep-seated cares and superficial or peripheral cares in order to allow patterns to constitute a species of cares. See Blustein (1991, 46) for the distinction between deep-seated caring and peripheral caring.
5.2 Cares without patterns

The second counterexample that I will introduce here is, I think, even more forceful than the previous ones. It aims to show that one can care about something even though one does not instantiate the pattern required by the pattern-based account.

The counterexamples involve people who, although they may deeply care about a particular thing, fail to feel the relevant emotions on certain occasions. Of course, the pattern theorist can allow occasional gaps in patterns (Helm, 2001); as long as they are only occasional, that is, as long as they do not falsify the claim that there is a pattern, the subject can still be said to care. The pattern-based account, as a result, requires that cases of cares without patterns be impossible.

Before considering a case, let me make two assumptions. First, I assume that the absence of a pattern in the actual world implies, on the pattern-based account, the absence of a care. Although part of the pattern may be counterfactual only, there must be a part of it that is actual, and that part must constitute itself a pattern—it cannot be the case that one feels one emotion towards something in the actual world but would feel all the rest of the pattern in another world. This is to rule out, for instance, the attribution of a care for me to everybody, since there is presumably a world at which every single person on this planet instantiates a pattern of emotions whose focus is me. Second, I assume a certain view on what it takes on the pattern-based account to have a pattern that is sufficient for caring. According to Helm, we have seen, collections of emotions should satisfy, at least in most cases, the two rational requirements introduced

\[31\] At least insofar as the relevant circumstances have occurred in the past. Perhaps the pattern theorist could allow cares without patterns of actual emotions when the situations relevant to the triggering of the emotions have never shown themselves. For instance, one could care for one’s life even though one has always lived in an extremely safe environment, and as a result has never faced any danger. Whether the pattern theorist would allow this kind of case, and how she would account for it, is a matter of debate that need not concern us here. Even though the pattern theorist may allow cares without patterns of actual emotions in cases where none of the situations relevant to the care have been confronted, she surely would not allow cases where no such pattern is present and the relevant situations have been confronted. Thanks to Moritz Mueller and Michael Scott for pressing me on this point.
above. For instance, when I feel hope at $t_1$, I am committed to feeling (say) relief, sadness, or frustration at a later time $t_2$ depending on what goes on at that time. Should I repeatedly fail to feel the emotions that earlier emotions committed me to, I would not count as genuinely caring. On this account, therefore, I won’t count as instantiating a pattern, and therefore as caring, if I only undergo two emotional episodes with a common focus over a ten-year period in which I have had the chance to confront many situations relevant to the care in question. What we need to know, now, is whether it is possible to plausibly attribute a care to someone who does not instantiate a pattern with a certain focus and whose members are rationally interconnected.

I think it is. Consider the following case (adapted from Goldie, 2011). John is both a wonderful husband and a wonderful father. Although he has a stressful job, he always manages to give an important place for his family in his life. Above all, he feels joy every time they are around him, he misses them when they are not there, and he worries when he hasn’t heard from them in a while. All the emotions that one can think of that a loving and caring father would have, John has them. After his mother died, John begins to be strongly depressed, and both his behaviour and emotions in turn change radically. John doesn’t behave and feel the way he used to. And nothing in his behaviour shows evidence that he still cares for and loves his family. Yet, even though we don’t have evidence for this, the intuition goes, it would be harsh to deny that John may still, ‘deep down’, care for his family.

Now, imagine that, after a couple of years of depression, the father quickly comes back to his initial state of mind. He cares about and loves his family just as much as before, with the same regularities in behaviour and feeling. There are two main ways that the pattern theorist could interpret this case. First, in line with the pattern-based account, the before-depression period involves a certain pattern and the after-depression period involves a similar but distinct pattern. A problem for this interpretation of the case is that it doesn’t explain why the pattern involved in the before-depression period is very similar, perhaps exactly similar, to the pattern involved in the after-depression period. The second
interpretation, also compatible with the pattern-based account, is that the before-depression pattern and the after-depression pattern are themselves part of larger pattern which has a gap that is not significant enough (at least relative to the life of the subject as a whole) to threaten its identity. Presumably, therefore, John could be counted as still caring for his family while he was depressed, a similar way one can still count as caring for one’s family when one is tired and insensitive one evening. The problem, however, is that John could have failed to recover from his depression, or could have died before recovering. So, if the pattern theorist accepts the claim that John could still care for his family during his depression, this implies that caring here is hostage to fortune: whether or not John still cares for his family at a time during his depression depends on how he would feel and behave at a later time. Intuitively, however, whether or not a subject cares about something at a time does not depend on future facts about her. At best, the future facts constitute a way to test which state of affairs obtained.

Another problem with this interpretation of the case is that it is not clear why the gap must not be significant in order for a care to be had. Why should the gap be small compared to the subject’s entire life? Why can’t we have a very big gap without a pattern and still have the relevant care? Presumably, it is because the bigger the gap is, the less justified we are in thinking that there is a pattern in the first place. The problem, therefore, is why we can’t attribute a care in the absence of a clearly defined pattern. Imagine now that John sadly does not recover from his depression, and that he has it for the rest of his life. Imagine now that over this very long period, John has random and very rare moments where his depression subsides and he feels caring emotions for his family, such as joy and pride at the view of his grown-up and successful daughter. Assuming that the relevant emotions are too distant in time from each other, and insufficiently systematic, to form a pattern, we nonetheless do not wish to rule out the possibility that John still cares for his family.

The problem can be rephrased as a problem about explanation. Intuitively, the explanation why John does not feel anything for his family is that he is
depressed. This much is something the pattern theorist can accept. The explanation that the pattern theorist is unable to give, however, is one about the role that the depression plays here and about why John feels random bouts of emotion for his family over time. According to the pattern theorist, the role of the depression in the last case is to remove (or destroy) one’s cares, given that the depression removes the relevant patterns. This explanation, however, does not seem to have the resources to explain why John starts to feel emotions, as opposed to nothing, for his family as soon as his depression is less strong (which, we have assumed, does not happen often). If John’s care for his family disappeared as a result of the depression, we wouldn’t expect John to feel for them as soon as the depression subsides. We would rather think that John would need much more time to get back to his initial state. An alternative explanation, in contrast, is that the role of the depression is to block John’s cares, which in turn accounts for why John may feel certain emotions whenever the depression is less prevalent; the fact that the depression goes away, even for a moment, gives John’s heart a chance to express itself (Goldie, 2011).

The foregoing reasoning suggests that patterns of rationally connected emotions are not necessary for caring. The pattern theorist does not have the resources to explain the relevant failures to feel other than that the subject doesn’t really care about the object, and thereby fails to do justice to a commitment we have to the effect that, in certain conditions, people’s failures to feel is not decisive evidence that they don’t care, and that there is a special explanation for such cases, as in the depressed father case.

6. The Simple View without patterns

So, the claim according to which the instantiation of a care is the instantiation of a pattern of emotions, at least understood as a conceptual claim, is at best inconclusive. The Simple View need not lead to the pattern-based account, however. Recall that, according to the Simple View, to care is to be disposed to feel in some ways on some occasions. The pattern theorist, we saw, understands the locution ‘to be disposed to’ conditionally; according to her, X is disposed to
feel in C just in case X would feel in C. Put in terms of patterns, X is disposed to feel in C just in case there is a pattern of actual and possible events where X feels in C.

Another way to put the reason why the pattern-based view should be rejected is that, while cares are intuitively deeply embedded in one’s psychology, there is nothing in the idea of a pattern of emotions that makes it necessarily embedded in this intuitive sense. On such a view, indeed, cares at a time depend on things at other times/worlds. As I will later argue, cares should rather be conceived as properties of persons that are intrinsic to them and whose identity at a time does not depend on what happens at other times.

In the following chapters, I will argue that the mistake of the pattern theorist lies in the way she construes the notion of a sentiment. On an alternative view, sentiments are distinct from—though conceptually connected to—their manifestations, and at least many of them are intrinsic to their bearers. Such realist, non-reductive account of emotional dispositions will be developed in Chapter 3. This account, we will see, is motivated by a realist, non-reductive view about dispositions in general (Chapter 2). Given that caring is a kind of sentiment, my account of sentiments will imply a realist construal of the Simple View.

In addition, the intuitive thought that cares entertain a close relationship with patterns of emotion will be accommodated, as patterns will be conceived as more or less reliably produced effects of cares, and therefore as (defeasible) evidence for them; and so what distinguishes patterns that are not expressive of cares from those that are is to be found in their dispositional origin. In an important sense, therefore, whole patterns of emotional dispositions characteristic of caring are explained by cares, contrasting with the pattern-based account which could only explain particular instances of emotion by appealing to patterns (and thereby to cares).
7. Conclusion

I started out this thesis by noticing the relative neglect of the notion of sentiment in contemporary philosophy of emotion. From the title of this thesis, we can also guess that my aim here is to provide a ‘defence’ of sentiments. It may be wondered then why I have so far chosen to focus exclusively on caring.

If the argument in this and the next chapters is on the right track, caring turns out to be a species of sentiment construed as a distinct kind of entity from emotional episodes but nonetheless related to them in some way (the latter being, as we will see, the sentiments’ ‘manifestations’). So, if something as important to us as caring should be conceived as falling under the neglected category of sentiments, I take it that this is strong reason to start being interested in sentiments. Of course, having a plausible account of caring is important in its own right. However, it should be kept in mind that the development of this account can also be seen as an argumentative strategy, a means to a certain end—that of vindicating the category of sentiments as a suitable object of inquiry distinct from the category of emotions—and therefore that it can have far-reaching implications.

The next two chapters will be dedicated to, first, showing that dispositions in general should not be reduced to events, and, second, showing that positing emotional dispositions, construed as irreducible to mental events (such as emotional episodes), allows us to explain our emotional lives in a more satisfactory way.
Chapter Two

The Nature of Dispositions: From Semantics to Metaphysics

1. Introduction

As seen in the previous chapter, a distinction is sometimes made in the philosophical literature on emotion between emotions as occurrent or episodic states—what I simply called ‘emotions’ or ‘emotional episodes’—and emotions as dispositional states—what I called ‘sentiments’. An important task is to know how best to construe this distinction. The question is, what is the relationship between ‘occurrent’ emotions and ‘dispositional’ emotions? A way into the discussion is to ask the following question: what is the relationship between emotion locutions that seem to refer to occurrent emotions and those that seem to refer to dispositional emotions? When I say that John is angry with Sarah, what is it that I am saying? Am I saying that sometimes John has a bout of anger towards Sarah? It seems that I need not be saying this. I could be saying that John is disposed to feel anger towards Sarah in certain situations, for instance when she reminds him of what made him angry in the first place. But now, the question is, what kind of thing is ‘being disposed to feel e’, if it is a kind of thing at all? Is it a real property that people can have and share? Or are such locutions supposed to refer to no specific property? One possibility is that such locutions are to be analysed as shorthand for a conditional statement about how the relevant person would behave in a range of counterfactual situations. On such a view, John is angry (in the dispositional sense) with Jane just in case, if he were to be in situations $s_1, s_2, \ldots, s_n$, he would, among other things, get angry. I will argue that such a view is mistaken, both about dispositions in general and

---

32 In this chapter, I take ‘occurrence’ to be synonymous with ‘event’ and ‘episode’. See Ryle, 1949, for the distinction between dispositions and episodes.
emotional dispositions (or sentiments) in particular. Rather, emotional dispositions are to be construed as properties that explain the occurrence of the stream of emotional episodes that we experience from time to time.

In the following, I will ask the question of how best to construe the notion of an emotional disposition at play in common-sense and what work, if any, it can perform in an explanation of our emotional lives. The question that will concern us, therefore, is how best to construe the notion of an emotional disposition. I will consider two main possibilities:

(1) **Realism**: Emotional dispositions are psychological properties in their own right, existing over and above the events (e.g., emotional episodes, behavioural effects, etc.), actual or possible, that they typically manifest.

(2) **Antirealism**: Emotional dispositions are nothing over and above their manifestations.

An antirealist construal of discourse about sentiments is understandable, as it does not commit us to the existence of an extra entity in our ontology of the mind; it doesn’t require one to show that there is more to our emotional life than sequences or patterns of discrete emotional episodes (or, at the very least, it allows one to remain silent on the further question whether there is more than that). By contrast, a realist construal (by which I mean a construal that takes at face value sentences whose surface grammar quantifies over dispositional properties) that is not motivated can seem either as merely apparent or in ‘need of redemption’ given the alleged ‘ethereal’ character of dispositions (Goodman, 1983, 40). Given the controversial nature of the notion of a disposition, a proper defence of realism about emotional dispositions is therefore needed.

This defence will be in two parts. First, I will show that much of our dispositional discourse and concepts in general is based on realist assumptions,
thereby rejecting the antirealist’s interpretation of the discourse and suggesting that realism about emotional dispositions is not a revisionary position. This will be the main task of this chapter. Second, I will argue that at least some dispositional properties are explanatory (later in this chapter) and that emotional dispositions or sentiments are among them (Chapter 3). As a result, a common suspicion about the explanatory potential of dispositional properties will be assuaged. I will argue that a proper realist account of emotional dispositions can do more than prove them innocent; it can provide insights into our emotional lives that cannot be provided by the antirealist alternative. The aim of this chapter and the next is not only to defend the reality of emotional dispositions (or sentiments), but also to defend their centrality.

2. The problem with dispositions

Although an important part of the way we ordinarily think about the world, dispositional concepts are the kinds of things that appear elusive and hard to explicate once we try to think about them. What is it for a glass to be fragile, for a sugar cube to be soluble, for a person to be irritable? An intuitive way to make sense of these ascriptions is to say that dispositional predicates are aimed at capturing how the object (or person) would behave in various circumstances; they are, so to speak, about the possible, rather than the actual: the glass is fragile in virtue of the fact that it would break in certain (non-actual) circumstances, the sugar cube is soluble in virtue of the fact that it would dissolve in certain (non-actual) circumstances, and so on. Thus, if dispositions are to be construed as instantiated properties, as they seem to be—a unbroken glass or vase can be fragile at any moment of its existence—and therefore as actual, they may appear too mysterious, or in Goodman’s terms, ethereal (Goodman, 1983, 40), to be accepted in one’s ontology along with more legitimate properties such as shape and size.

Such a tension—between the apparent property-status of dispositions and their being about possible states of affairs—can be resolved in several ways. One standard solution, however, is to deny the status of properties to dispositions
altogether, and claim that disposition ascriptions commit us to nothing more than how the objects to which the dispositions are ascribed would behave in certain circumstances. More precisely, rather than claiming that disposition ascriptions commit us to the existence of both a dispositional property and the events that are its manifestations, the strategy aims to show that such ascriptions commit us to events only. Nothing like unactualised *possibilia* or ‘potential beings’, lurking “in a mysterious realm intermediate between potentiality and actuality” (Mumford, 1998, 4), are needed. Dating back to David Hume (1955), this position has attracted many writers in the past decades, and, although increasingly less popular today, is still assumed to be true in much contemporary work outside research on dispositions.

In the following, I will argue that the Humean, antirealist view of dispositions is mistaken, for, first, the semantics of disposition ascriptions it is associated with suffers from serious problems and, second, dispositions as properties can play important explanatory roles that the Humean cannot allow. It is only if dispositions are to be construed as *actual* properties, ontologically distinct from their possible manifestations, that a great deal of common-sense talk and thinking can be preserved, and thereby that the tension mentioned earlier can be resolved. According to the particular kind of realism about dispositions I will put forward, dispositions and manifestations are, *contra* Hume, distinct entities. And this fact will be the one to do the main explanatory work when we move on to the topic of emotional dispositions.

3. Preliminary distinctions

Before tackling the question of how best to understand the notion of a disposition, let’s introduce two distinctions that will prove important at various stages of the discussion. The first has to do with the kind of realism defended in this chapter and the specific views it is contrasted with; the second has to do with the methodology involved in the analysis of disposition ascriptions.
3.1 Realism and antirealism about kinds of properties

One thing to notice is that one can be a realist (as I understand the term in this chapter) about dispositional properties and think that dispositional properties are not part of the ‘ultimate constituents’ of reality, a similar way one can be a realist about persons and think that the property of being a person is ultimately reducible to non-personal properties. Our acceptance of the fact that dispositions may be reducible to other things, then, does not automatically make us antirealists. Yet, we have seen that the antirealist performs some sort of reduction of dispositional properties by reducing talk of them to talk about events. Why does the antirealist not count as a realist, then, if reduction per se does not imply antirealism?

The answer has to do with the kind of thing dispositions are reduced to. The kind of reduction the realist allows is one that grants dispositions the status of properties even after they have been reduced. If A is a dispositional property, and B is a non-dispositional property, and if A = B, then all the explanatory work we thought A does is work that is done by B. In such a case, reduction preserves the explanatory power of dispositions. By contrast, the antirealist does not preserve the thought that dispositions are properties that enjoy some degree of explanatory power qua properties; instead, she claims that all the work we thought dispositions could do is actually done by another kind of entity, namely events. Talk about dispositional properties is, on her view, just another way of talking about such events. This chapter will be primarily dedicated to criticising this semantic thesis.

---

33 David Armstrong (1968), for instance, claims that dispositional properties are grounded in non-dispositional properties (their ‘categorical’ base) such as properties about microphysical structure, shape, size, and so on, while holding the claim that antirealism cannot account for the semantics of disposition ascriptions.

34 Namely things where properties are instantiated but which are not exhausted by these properties.
To be sure, not all antirealists about dispositional properties make such an analytic reduction. One could indeed agree that the semantics of disposition ascriptions is best described in the realist’s terms, yet deny that disposition terms pick out anything in the world. One would thereby be an error theorist about dispositional discourse, claiming that whenever we speak about things such as fragility, irritability, perhaps even love, there is nothing in the world corresponding to the relevant terms, and thereby that what we say in the relevant contexts is systematically untrue. Why would the error theorist think that there are no genuine dispositional properties in the world, as the realist thinks of them? There are many possible reasons, but one that I will later focus on is the purported fact that dispositions lack genuine explanatory power. It is not just that dispositions explain phenomena that could be explained by properties that are identical with them (realism, as we have seen, is compatible with this possibility), but that dispositions, as the realist thinks of them, compete with non-dispositional entities in the explanation of certain phenomena and fail; there is no work left to the dispositions.

Let’s call the antirealism based on the reduction of talk of dispositions to talk of events ‘reductionism’ and the antirealism based on the claim that dispositional properties as the realist construes them do not exist ‘error theory’. While the reductionist will typically want to preserve dispositional discourse provided that the reducing discourse is acceptable to her, the error theorist will have a reason to eschew the discourse altogether, thereby deciding to opt for eliminativism about it, unless she thinks that the claims of the discourse are nonetheless worth uttering in certain contexts (adopting in turn of form of fictionalism).

Antirealism, of both stripes, should be conceived as global theses, namely as theses aiming at capturing the whole truth about dispositions. Reductionism is the thesis that all—not just some—disposition ascriptions can be paraphrased into conditional statements about events involving the object of the ascription. By contrast, error theory is the thesis that, although disposition ascriptions

35 The pattern theorist from Chapter 1 counts as a reductionist.
should be taken at face value (i.e., as purporting to refer to genuine dispositional properties), no dispositional properties exist. The realist, by contrast, denies the reductionist’s claim that all disposition ascriptions can be paraphrased in the way she suggests, and reject the error theorist’s claim that dispositional properties do not exist.

Notice though that realism as defined here is compatible with the claim that some disposition ascriptions can be paraphrased in the way the reductionist suggests; as I will suggest later in Section 10, a certain class of disposition ascriptions may best be captured in reductionist terms. The realist, moreover, can accept that some disposition ascriptions purporting to refer to genuine dispositional properties can fail in that aim. The realist, in other words, may only make the modest—and sensible—claim that many dispositional properties that we can think of exist, not that they all do.

In the following, I will argue against both forms of antirealism, first by showing that many disposition ascriptions are best interpreted in realist terms, and second by showing that at least some dispositional properties are explanatory. It will turn out that many paradigmatic disposition ascriptions seem to genuinely refer to dispositional properties.

3.2 Covert/conventional and overt/canonical disposition ascriptions

Most everyday names for dispositions do not explicitly mention the dispositional character of the things they are applied to. When we say that a glass is fragile, we are not explicitly saying that the glass is disposed to break in certain circumstances. Nevertheless, this is what we think we mean when we call a glass fragile. What we mean, it seems, and even though we don’t say it, is that the glass has a certain disposition to behave in a certain way in various situations. Let’s call ‘covert’ disposition locutions, such a ‘being fragile’, that are only implicitly dispositional, and ‘overt’ disposition locutions, such as
‘being disposed to break when struck’, that are explicitly dispositional (Bird, 2007, 18-19, Choi, 2008; see also Manley & Wasserman, 2008).

An analysis of disposition ascriptions thus involves two steps (Lewis, 1997, Choi, 2008). First, we translate covert disposition locutions into overt ones. ‘Being fragile’ is defined as ‘being disposed to break when struck’. Then we attempt to give an analysis of overt disposition ascriptions:

\[ X \text{ is disposed to break when struck if and only if...} \]

The payoff of translating covert locutions into overt locutions is that a canonical formulation of overt disposition ascriptions can be given. The task of the second step can therefore be the one of giving an analysis of the following general form:

\[ \text{An object } o \text{ is disposed to give response } r \text{ to stimulus } s \text{ if and only if...} \]

Much of the contemporary debate over the nature of dispositions is concerned with providing an adequate analysis of this canonical formulation. It is important, though, to keep in mind the fact that an analysis of disposition ascriptions requires the preliminary work that the first step involves. As we will see in Section 7, the question of how to make the transition from covert to overt locutions may be crucial in assessing various proposals of the analysis of the canonical disposition ascription.

4. The simple conditional analysis

If you asked me what it is for a glass to be fragile, I would spontaneously respond that it is for it to be disposed to break when dropped or struck. It is, it
seems, what it *means* to be fragile. But then, you could ask me, what is it to be disposed to break when struck? At least if I’m not in a reflective mode, I would probably respond that for an object to be disposed to break when struck is for the following counterfactual to be true:

*If the object were struck, it would break.*

If my response to your query is to be taken at face value, a disposition ascription such as ‘The glass is fragile’ can be paraphrased without loss of meaning into a certain counterfactual conditional: all it takes for the disposition ascription to be true is that the object in question would behave in certain ways in certain circumstances.

Generalising, we get the following analysis of (canonical) disposition ascriptions:

Simple conditional analysis (SCA):

*Something* $x$ *is disposed at time* $t$ *to give response* $r$ *to stimulus* $s$ *if and only if, if* $x$ *were to undergo stimulus* $s$ *at time* $t$, *$x$ would give response* $r$.\(^{36}\)

(SCA) has been defended explicitly by Ryle (1949), Goodman (1983), and Quine (1960), and has been assumed to be true by countless others.

Besides its seemingly intuitive plausibility, (SCA) has the advantage of not committing us to the existence of anything more than events. On this distinctly Humean picture, dispositions are not ethereal, occult entities, hiding behind the events we observe, as there is no genuine distinction to be drawn between a power and its exercise. And this is a fact that our discourse about dispositions is

---

\(^{36}\) This is Lewis’ formulation (Lewis, 1997, 143).
consistent with. The Humean picture is in great part motivated by an empiricist epistemology. Stephen Mumford nicely captures the Humean picture in the following:

If we are to be consistent empiricists, then all we ought to admit of is observation of such events. This is all that there is available to which an ascription of a power can refer. We have acquaintance with nothing else, hence the true reference of a disposition ascription is a set of events that are expected by us, given the evidence of previous events. The parallel with Hume’s reductive account of causal relations is obvious. Our belief in real causal connections between events has no logical justification; it arises merely from custom and habit. (Mumford, 1998, 36-37)

On the Humean picture, there is no room for dispositions construed as properties ‘over and above’ the events that we observe or could observe on certain occasions, and therefore as properties that can vary (e.g., be acquired and lost) independently of such events.

5. Counterexamples to SCA: Finks, masks and mimics

(SCA) is subject to a number of counterexamples. The upshot of these counterexamples will be that the apparent reference in disposition ascriptions to dispositional properties independent of their manifestations cannot be analysed away.

Some counterexamples are such that an object can have a disposition while the conditional associated with the disposition is false (i.e., counterexamples to the left-to-right implication of the analysis), while some counterexamples are such
that the conditional associated with a certain kind of disposition is true of an object while the object lacks the disposition (i.e., counterexamples to the right-to-left implication of the analysis).

C.B. Martin (1994) provides cases where a disposition is ‘finkish’. In its initial version—let’s call it \( F1 \)—an object \( O \) lacks a certain disposition \( D \), but whenever \( O \) finds itself in the conditions \( S \) sufficient for \( D \), were it possessed by \( O \), to be manifested, \( O \) acquires \( D \). Consider the case of a wire that, although lacking, at any given time when it is not touched by a conductor, the disposition to go live when touched by a conductor, acquires it every time it is touched by a conductor. In such a case, the left-hand side

\[
(1) \, \text{The wire is live}
\]

of the bi-conditional is false, while the right-hand side

\[
(2) \, \text{If the wire were touched by a conductor, then electrical current would flow from the wire to the conductor}
\]

is true.

Another finkish case—call it \( F2 \)—is when an object \( O \) possesses a disposition \( D \) but that, whenever \( O \) finds itself in the conditions \( S \) sufficient for the manifestation of \( D \), \( O \) loses \( D \). Consider now the case of a wire that, although live at any given time it is touched by a conductor, becomes dead every time it is touched by a conductor. In such a case, the antecedent (1) is true while the consequent (2) is false.
A second kind of case involves objects that manifest the events associated with a certain kind of disposition but that are not in any way responsible for the triggering of these events (Johnston, 1992, A.D. Smith, 1977). In such cases, the object merely ‘mimics’ the manifestation of a disposition, given that no disposition of the object is responsible for the relevant events. Mimicking cases are cases where no change occurs in the object, while finking cases are cases where an object changes in the dispositions it has—either by having one of its dispositions removed or by acquiring a new disposition.

Two variants of mimicking cases are possible. First, an object can lack a disposition D but behave as though it had it (call it MI1). Imagine the case of a person who, although not shy—in the sense of being disposed not to talk much at social gatherings, say—is paid to act shyly. In such a case, the statement ‘X is shy’ is false while the statement ‘if X were around people, he would not talk to them a lot’ is true.

The second kind of mimicking cases has it that an object can have a given disposition D, behave as though it possessed D, while the behaviour in question is not a manifestation of D but of something else; let’s call it ‘veridical mimicking’ (or MI2). Now, we can imagine a case where a shy person is hired to act shyly (i.e., not to talk much when around people). Although this does not constitute a direct counterexample to (SCA), this case seems to suggest that the link between a disposition and its manifestation must be robust enough so as to rule out cases where the conditional is true only accidentally.

Masking cases—MA—are, like mimicking cases, cases where no change seems to occur in the object; no disposition is either removed from or acquired by the object. But they differ from mimicking cases in that no manifestation is displayed. In such cases (due to Johnston, 1992 and Bird, 1998), an object is disposed to exhibit a given response in certain circumstances but fails to do so because of some interfering factor preventing that manifestation from occurring.
by blocking the process supposed to obtain between the stimulus and the response. So, we have here a true disposition ascription (‘X is disposed to P when C’) whose associated conditional (‘If C were the case, X would P’) is false. Consider the case of a fragile glass that is wrapped up in some packing material and that as a result doesn’t break when struck. Assuming that being fragile is being disposed to break when struck, we can say that the glass is disposed to break when struck but wouldn’t break if struck. As a result, although the following disposition ascription

*The glass is disposed to break when struck* (or: *The glass is fragile*)

is true, the following conditional

*If the glass were struck, it would break*

is false.

Alexander Bird (1998) came up with an interesting case of masking where an object has its disposition masked after it has received the stimulus required for the disposition to manifest itself—he calls this kind of case ‘antidote’. Consider the case of someone who, immediately after having ingested a poison (which is disposed to kill when ingested), is administered an antidote that breaks the causal chain between the ingestion of the poison and the death. It is important to notice that the difference between the finkish case F2 (i.e., the case where an object has a disposition removed) and the antidote case is to be found in what is going on in the time gap between the stimulus and the expected response. In F2, an object disposed to give response \( r \) to stimulus \( s \) receives \( s \) at a time \( t_1 \) and, while on (SCA) it is supposed to give \( r \) at \( t_3 \), doesn’t because it loses its disposition at \( t_2 \). In the antidote case, by contrast, no loss of disposition occurs
between \( t_1 \) and \( t_3 \); the object’s disposition is rather prevented from manifesting itself. The disposition is merely held in check, not eliminated (Bird, 1998, 229).

Together, finkish, mimicking and masking cases appear to give us a strong case against (SCA).

6. Lewis’ reformed conditional analysis

A seemingly obvious way to explain the intuitive force of the counterexamples is to say that the original analysis fails to appreciate the fact that dispositions are an “intrinsic matter” (Lewis, 1997). Indeed, as Molnar says, “[t]he conditional does not seem to be saying anything about what it is, in the object, that makes the response follow the stimulus.” (Molnar, 1999, 1) We have seen that (SCA) faces three main problem cases:

1) Cases where the antecedent of the analysis is false but the consequent is true (F1 and MI1)
2) Cases where the antecedent is true but the consequent is false (F2 and MA)
3) Cases where, although both the antecedent and the consequent are true, this is a fluke (MI2)

Rather than denying the force of these cases, and saving in turn (SCA), some people have attempted to modify the analysis accordingly. One popular approach is the one developed by Lewis (1997). Focusing on finkish cases, Lewis modifies (SCA) in a way that does justice to the rather natural thought according to which the removal and acquisition of dispositions require some intrinsic change. Adopting the two-step approach exposed above, he gives the following analysis of canonical disposition ascriptions (Lewis, 1997, 157):
Reformed conditional analysis (RCA):

*Something x is disposed at time t to give response r to stimulus s if and only if, for some intrinsic property B that x has at t, for some time t’ after t, if x were to undergo stimulus s at time t and retain property B until t’, s and x’s having of B would jointly be an x-complete cause of x’s giving response r.*

Where an ‘x-complete cause’ is “a cause complete in so far as havings of properties intrinsic to x are concerned, though perhaps omitting some events extrinsic to x” (1997, 156) Note that the intrinsic property B that Lewis mentions in the analysis need not be itself a dispositional property; rather, it is construed as a property in virtue of which something can be said to have a disposition. Whether or not such a property is, metaphysically speaking, essentially dispositional (whatever that means) is open to debate. In the following, I will however call ‘dispositional property’ any intrinsic property that does the work suggested by (RCA), putting aside the subtle issue of whether properties can be said to be essentially powerful.

(RCA) easily accommodates cases of finks. In the first finkish case, the wire cannot be said to be live at the time of ascription (i.e., when not touched by a conductor) because it does not have at that time the intrinsic property required by the analysis. In the second case, by contrast, the wire can be said to be live at the time of ascription because, at that time, it possesses the required intrinsic property, even though it will lose it when touched by a conductor (for it doesn’t retain it for a sufficient amount of time for the manifestation to occur).

While (RCA) accommodates cases of finks, it isn’t clear whether it accommodates cases of masks and mimics. On the one hand, although dispositions are intrinsic on Lewis’ view and masks are supposed to be extrinsic, masking cases seem to be genuine counterexamples to (RCA), at least taken as including an implicit universal quantification (being prefixed by an
‘For all times t’ clause\(^{37}\). It is certainly quite plausible that a fragile glass wrapped in some packing material retain its dispositional property, and that even though it wouldn’t break if it were struck. On the other hand, mimicking cases are not ruled out by the analysis, but do not obviously pose a problem to it.

Now, Lewis’ reaction is to attempt to debunk the apparent force of these counterexamples. First, as a way to counter mimicking cases, Lewis denies that “dispositional concepts generally have built-in response-specifications requiring a direct and standard process.” (154) By this he means that, although some dispositional concepts require a certain kind of process between stimulus and response, not all do. He contrasts fragility with the HIV virus. Assuming that being fragile is being disposed to break when struck, consider the case of a Styrofoam dish that makes a distinctive sound every time it is struck. Now, imagine that the Hater of Styrofoam, whenever he hears this sound, “comes and tears the dish apart by brute force.” (153) As a result, Lewis says, Styrofoam dishes are disposed to break when struck whenever the Hater is within earshot. Are such dishes fragile? It makes sense to answer negatively, and claim that “there is a certain direct and standard process whereby fragile things most often (...) break when struck, and the styrofoam dishes in the story are not at all disposed to undergo that process.” (ibid.) Taking being lethal for a virus to be defined as being disposed to kill, by contrast, Lewis gives the example of a virus that is “disposed to cause those who become infected with it end up dead before their time, but not undergo the direct and standard process whereby lethal viruses mostly kill their victims.” (154) The HIV virus, unlike other ordinary viruses, does its job, not by attacking the subject, but by weakening her immune system. “For this virus does not itself interfere with any of the processes that constitute life. Rather, it interferes with the victim’s defences against other pathogens” (ibid.). Still, the HIV virus is obviously lethal. Given that dispositional concepts in general are not process-specific, it seems therefore that (RCA) is not weakened by cases of mimics.

\(^{37}\) Molnar, 1999, 7.
Lewis’ response to mimicking cases is insufficient, however, given his analysis. The important thing to note here is that an adequate analysis is supposed to yield the right results when it comes to particular disposition ascriptions. If, for instance, we have a case where a glass is definitely fragile but where the relevant *analysans* (e.g., a conditional statement) is not true of it, then we have a counterexample to, and therefore a possible refutation of, the analysis. Lewis’ analysis, unfortunately, falls prey to a possibility he wanted to allow. As Lewis admits, the Styrofoam dish case shows that fragility *does* require a certain kind of process. However, as Molnar notes, as it stands (RCA) “does not merely *permit* process-unspecific dispositions, it ensures by definition that every disposition *has to be* process-unspecific.” (Molnar, 1999, 5) All we need to do is therefore to find a case where we have the stimulus, intrinsic property and response corresponding to a certain kind of disposition, and where the intrinsic property is the intrinsic cause of the response, but where the response cannot be said to be the manifestation of the disposition. Molnar imagines the case of a virus that is developing in a land where a cautious tyrant has decreed that “whenever a hitherto unknown virus infects someone, the infected person is to be killed forthwith for the protection of the rest.” (5) In such a case, we have the stimulus (the entry of the virus), the intrinsic property (the hosting of the virus) and the manifestation (death) that all seem to satisfy the terms of the *analysans* of (RCA). Moreover, the hosting of the virus is arguably a cause of the death. Nonetheless, the virus could well be harmless, i.e. not lethal. We therefore seem to have here a counterexample to (RCA).

Lewis’ strategy to weaken the apparent force of masking cases is to deny that they constitute a counterexample to (RCA). Rather, such cases show the deficiency of the definitions we give of covert disposition locutions. Recall that Lewis’ analysis requires us to work out the meaning of ordinary disposition ascriptions in order to put them in overt forms. We have seen that *being fragile* can be defined as *being disposed to break when struck*. Such preliminary work, Lewis notes, is obviously sketchy, as the definitions we give are typically loose ones. Masking cases, according to Lewis, are cases where we can see the
shortcoming of such definitions. Take the case of the glass wrapped up in some packing material. It is strongly intuitive to think that the glass is nonetheless fragile. Lewis doesn’t deny that this may be true, but still thinks that (RCA) is not threatened by it. All (RCA) says is that the glass is not *disposed to break when struck*, not that it is not *fragile*. It is only if we think that ‘being disposed to break when struck’ is *synonymous* with ‘being fragile’ that we can take the case to be a counterexample to (RCA). However, such a claim about the meaning of *being fragile* can be doubted. *Being fragile*, it may seem, is *more* than *being disposed to break when struck*. It is, perhaps, *being disposed to break when struck and not wrapped in some material that would prevent breakage*. When it comes to the characterisation of everyday dispositional locutions, Lewis says, “the specification both of the response and of the stimulus stand in need of various corrections.” (1997, 153) To take another example, we can roughly define *being poisonous* by *being disposed to kill when ingested*. However, as it is acknowledged, this is only rough; it is only if an *antidote* is not ingested that a substance can be poisonous, yielding the following characterisation of *being poisonous*:

\[ X \text{ is disposed to kill when ingested and no antidote is taken.} \]

The clause in bold can be understood in two ways (Bird, 1998, 230-231). First, taking the case of *being poisonous*, we can understand the clause as stating a *pre-requisite* (or background condition) for a substance to count as poisonous. On such a reading, the poison’s disposition to kill when ingested is conditional upon the absence of an antidote. “The disposition is disposed to kill people who have not taken antidotes, but is not disposed to kill those who have.” (Bird, 1998, 230) As a result, a substance can be disposed to kill a person X at time \( t \), but not disposed to kill the same person at a later time because she has in the meantime taken an antidote. In such a case, the disposition of being poisonous comes out extrinsic to the poison—it depends on factors external to the substance in question—which is inconsistent with Lewis’ analysis.
The second reading is the one Lewis’ accepts. The exclusion of masks is part of the specification of the stimulus conditions. ‘Being poisonous’ is not defined as ‘being disposed to kill when ingested’, but rather as ‘being disposed to kill when ingested and no antidote is taken.’ If this is right, there is no contradiction in saying that a glass wrapped up in some packing material is both not disposed to break when struck and disposed to break when struck and unprotected by some packing material.

Such a response, however, is not without problems. I think there are easy problems and at least one hard problem, the latter being discussed later in this chapter. The easy problems are problems in response to which Lewis could, without much trouble, easily accept the consequences. The hard problem, by contrast, is not so easily put aside. First, the easy problems (all raised in Bird, 1998, 230-231). One consequence of Lewis’ strategy is that disposition concepts are far more complicated than we initially thought. “This means that the dispositions there are in the world are not the ones we think there are, because antidotes [or masks] are almost always possible.” (Bird, 1998, 231) Furthermore, there is a difficulty, perhaps impossibility, in spelling out all the possible interferers, leading to the conclusion that “[w]e could not properly characterize any real disposition” (ibid.). And even if we exclude the possible interferers only indirectly (e.g., ‘and nothing prevents the disposition from giving response r’), we will witness a proliferation of dispositions in the world. Even a brick, for instance, has a disposition to break when struck ‘and nothing prevents it from breaking’. (See next section for other problems with the strategy of excluding interfering factors en bloc.) This may be a problem if we assume that a view of dispositions should preserve the number of dispositions that we may pretheoretically think there are, an assumption Lewis is free to reject.
Now, the hard problem is simply this: why not apply the same strategy to finks too? If all counterexamples were dealt with by showing that they actually constitute a problem to the definitions we give of covert dispositional locutions, we wouldn’t need to modify (SCA) in the first place, and therefore (RCA) would be superfluous. This is the strategy employed by Sungho Choi (2008), which we will discuss in Section 7.2.

Before we move on, I would like to mention two other important issues that Lewis could not ignore; it is the possibility of extrinsic dispositions and the possibility of intrinsic finks, masks and mimics. On the one hand, if dispositions can be extrinsic, an analysis of disposition ascriptions in general that makes all dispositions intrinsic is inadequate. The possibility of extrinsic dispositions will be discussed in Section 8. On the other hand, if finks, for instance, can come in intrinsic versions, Lewis’ appeal to the intrinsicness of dispositions to get around these cases does not get off the ground. We will consider the possibility of such cases in Section 9.

### 7. Saving the analysis

In this section, I will discuss in more detail two variations on the strategy originating from Lewis, the strategy that Manley and Wasserman call ‘getting more specific’ (Manley & Wasserman, 2008). We have seen that a way to defuse some of the counterexamples to an analysis of disposition ascriptions is, to put it in general terms, to claim that we should be more specific in our characterisation of the stimulus and response associated with the disposition in question. This way of characterising the strategy is however neutral on precisely where the strategy is to be applied. Indeed, it seems that it could be applied either to the transition from covert to overt dispositional ascriptions or to the transition from overt dispositional ascriptions (i.e., the analysandum) to the analysans of the analysis. Although this distinction between two variants of the strategy doesn’t seem to have been made explicit in the literature, I think it is
worth making, for different problems can be raised depending on which variant we wish to pursue.

7.1 Getting specific: Modifying the analysans

The first strategy to consider is not Lewis’. It is not to modify the stimulus/response specification of the analysandum—the original characterisation of the latter is preserved. The strategy rather consists in modifying the stimulus/response of the analysans itself. To put it simply, covert dispositional locutions such as being fragile keep their initial definition; being fragile is being disposed to break when struck. But the analysis we give of the latter specifies the stimulus and response conditions in a way that rules out cases of finks, masks and mimics. The idea is to add a clause to the analysans of our preferred analysis that is designed to exclude possible interfering factors. I will use (SCA) as an example, but the strategy can be used by virtually anyone for whom finks, masks and mimics constitute a problem. Taking fragility as an example and adopting Lewis’ two-step approach, here’s how the strategy would run:

(A)(1) X is fragile ↔ (2) X is disposed to break when struck ↔ (3) If X were struck, then [restrictive clause] X would break.

The transition from (1) to (2) constitutes the required first step, and the transition from (2) to (3) constitutes the analysis we give to disposition ascriptions (here a modified version of (SCA)).

(A) constitutes a simple conditional analysis of the everyday disposition ascription of fragility except that it includes a clause whose function is to block possible cases of interference. You can here replace [restrictive clause] with clauses such as ‘ceteris paribus’, ‘in ideal conditions’, ‘in normal conditions’,

67
‘in most cases’, ‘in typical cases’, ‘in suitable cases’, ‘in a range of specific cases $c_1, \ldots, c_n$, or ‘in specific case C’. At least some of such clauses can in principle exclude all possible interfering factors, and therefore, it seems, save the analysis.

All these clauses, however, suffer from problems. First, the clauses appealing to ceteris paribus, ideal, typical, suitable, and normal or standard conditions, when spelled out completely, plausibly trivialise the analysis (Martin, 1994, 5-6). If such clauses are used, the worry goes, the analysis for fragility would ultimately boil down to:

*If X were struck, then, unless it doesn’t break, it would break*

which is vacuously true of everything that exists. This means, for instance, that bricks, stones, and any other thing that we intuitively would not call fragile, are fragile on this account, given that if they were struck, they would break unless they don’t. The challenge for the theorist wishing to appeal to the relevant clauses is to spell them out in a way that rules out all possible interferers without ending up with a trivial analysis, a task that has proved very difficult (though see Choi’s strategy in next sub-section).

Second, ‘statistical’ clauses such as ‘in most cases’ and ‘in a range of specific cases $c_1, \ldots, c_n$’ and ‘in specific case C’ seem to deny the possibility of dispositions that are always unmanifested, “by placing a priori constraints on the ratio of responses to stimuli…” (Molnar, 1999, 7)

Moreover, both proposals appealing to ranges of specific cases and single specific cases respectively fall prey to two counterexamples (Manley & Wasserman, 2008). Consider, first, the proposal that each disposition ascription is associated with a single specific stimulus condition (what Manley &
Wasserman call the ‘litmus test’ view, 2008, 67). A problem case for the resulting analysis is the case of Achilles’ heel. Consider a sturdy concrete block which, although not fragile, would break under exactly one condition C. If being disposed to break when struck is to be analysed in terms of condition C—an object is not fragile if it doesn’t break in that specific condition—then the brick counts as fragile, contra what we would ordinarily think. This cannot be right. “The block has an Achilles’ heel, but it will be (determinately) false in any ordinary context to say ‘it is fragile’ or ‘it is disposed to break when dropped’. It merely mimics fragility in a very specific circumstance.” (67) So the strategy of getting that specific fails.

Consider now the proposal that dispositional concepts are not associated with a single specific stimulus condition, but rather with a range of specific stimulus conditions. On such a view, a glass is fragile if and only if it would break in a range of specific situations. Manley and Wasserman imagine a case that is diametrically opposed to the case of Achilles’ heel, which they call reverse Achilles’ heel. Imagine the case of an extremely fragile glass which, by chance, “can withstand a surprisingly strong force, provided that the force is applied at exactly the right angle and at exactly the right point.” (69) Let’s suppose that the situation where the glass cannot break is paradigmatic to fragility. The problem is that, if fragility is to be analysed in terms of a range of specific stimulus conditions, conditions that include the one where the glass does not break, then it would force us to deny that the glass is fragile. This, again, can’t be right. So the strategy of getting specific by specifying a range of paradigmatic stimulus conditions fails.

According to Manley and Wasserman, the strategy of getting specific therefore fails, as “a disposition can be masked or mimicked even in conditions that are paradigmatic for the manifestation of that disposition—that is, conditions that will not be ruled out by getting more specific.” (66) This conclusion, however, may be premature, as we now need to consider the other variant of the strategy.
7.2 Getting specific: Modifying the *analysandum*

The second variant of the strategy of getting specific is the one adopted by Lewis in order to save his analysis against cases of masks. The strategy, we have seen, is to claim that masks show the shortcoming of the rough definitions we give of covert disposition locutions. A problem with Lewis’ approach, however, is its lack of uniformity in dealing with finkish cases, on the one hand, and masking cases, on the other hand (Choi, 2012). We have seen that Lewis deals with finkish cases by claiming that dispositions are intrinsic, and that he deals with masking cases by suggesting that what they show is that the rough translation of certain covert into overt dispositional locutions is inadequate. The question is simply, why can’t this move be made with respect to *finks* too? Choi (2008) suggests that Lewis’ strategy can be used in order to accommodate all problem cases, thus leaving open the possibility of adopting (SCA) as an analysis of canonical disposition ascriptions. The *analysans* is thereby kept intact; it is the *analysandum* that is modified, by specifying the stimulus conditions in a way that excludes the problematic factors. Applied to the case of fragility, the strategy looks like this:

(B) (1*) X is fragile ↔ (2*) X is disposed to break when struck [restrictive clause] ↔ (3*) If X were struck, X would break.

As we are dealing with specifying the stimulus conditions of *specific* disposition kinds, the restrictive clause need not be something like a ‘*ceteris paribus*’ clause as above. Instead, the clause can specify all the possible interfering factors *for the kind of disposition in question*. For instance, the specification of the stimulus conditions for fragility may require that the fragile object be unprotected by some packing material, suggesting in turn that the concept of fragility includes more specific stimulus conditions than we originally thought (Bird, 1998). If the restrictive clause specifies in some way all the possible interfering factors, including finks, then the analysis of (2*) needn’t be (RCA). (SCA), Choi
argues, can go through if all the problem cases have been carefully taken care of when specifying precisely the stimulus conditions of the disposition that we are trying to analyse.

One may object that it is just not possible to list all the possible interfering factors for any given disposition. However, the friend of the strategy can agree, as the interfering factors can be included indirectly, by appealing to ‘disposition-specific’ stimulus conditions and manifestation (Choi, 2008). The thought is that for each disposition name, and for each context of disposition ascription, there is an associated possibly rough-and-ready list of specific stimulus and manifestation conditions that somehow rules out problem cases of finks, masks and mimics. The existence of such a list is not difficult to establish, as we are capable of telling whether an object subjected to a fink or mask possesses the relevant disposition; there must be such a list, although it may be impossible to articulate it fully. Indeed, what drives our intuitions on various cases, including cases involving finks, masks and mimics, is plausibly an implicit acceptance of a set of conditions that must obtain in order for an object to have or lack of a given disposition. In addition, nothing prevents us from saying that the set of conditions may vary from one context of utterance to the next, as ‘fragile’ as applied to a glass may not mean the same thing as ‘fragile’ as applied to a house of cards. If Choi’s strategy is successful, (SCA) can thereby be preserved.

It also seems that Choi’s strategy does not obviously face the problem of Achilles’ heel, as the definition of disposition names include a range of stimulus conditions. A concrete brick that can break in only one specific context is not fragile, as it wouldn’t break in other situations that are specific or paradigmatic to fragility. Speaking in terms of paradigmatic situations may make us vulnerable to reverse Achilles’ heel cases though. Recall the case of the extremely fragile glass that does not break in a highly specific situation that is paradigmatic to fragility. We wanted to say that, though the glass should break in this situation, because the situation is paradigmatic to fragility, we do not
have the right to deny that it is not fragile because it doesn’t break. Now, if Choi’s strategy is to be pursued, for every way a disposition kind’s stimulus and manifestation conditions are going to be characterised, there will be a possible case where an instance of it will fail to manifest itself in one of the stimulus conditions. That is, there will be a case where the ‘disposition-specific’ conditions are going to be met but the relevant disposition is not going to manifest itself. This suggests, therefore, that it is in principle not possible to translate covert disposition locutions into overt, canonical ones in such a way as to block all cases of masks.

On top of that, other problems can be raised against Choi’s strategy. First, the strategy appears unable to account for the fact that some disposition kinds, such as fragility, are multiply realisable (Bird, 2007, 39). This, however, may not be a real problem for Choi, as he could well say that the precise meaning of ‘fragile’ varies from context to context, by being associated with different stimulus and manifestation conditions. Another problem is that covert/conventional dispositions are more explanatory powerful than any complicated and, as we have seen, incomplete overt counterparts (Bird, 2007, 40). Still, this is a consequence that Choi could accept.

Two problems are more serious, and concern Choi’s adoption of (SCA). We will see that it is sometimes hard to deny that at least some dispositions are an ‘intrinsic matter’. However, Choi’s adoption of (SCA) forces him to accept that all dispositions are extrinsic. For on (SCA), whether or not something has a disposition at a time when the disposition is not manifested depends on how it would behave at other times. Or, speaking in terms of possible worlds, it depends on how it behaves in nearby possible worlds where it is subjected to the relevant stimulus. For instance, a substance is poisonous—meaning: is disposed to kill when ingested and no antidote is taken—depends on how it would behave in situations where it is ingested and no antidote is taken. Consider a given substance that is poisonous but that is not now in the process of being ingested. If (SCA) is true, the substance’s being poisonous depends on how it
would behave in situations that are not the present situation. Since the fact that the substance is poisonous is not entirely determined by facts about the object at the present moment, being poisonous is extrinsic to the substance, which goes against a strong intuition of ours.

The second problem, which is in fact one that Choi himself raises (Choi, ms, 7-10, 2008, 45), is what he calls the ‘problem of random coincidence’ (ms, 7). On a standard Lewis/Stalnaker semantics for counterfactual conditionals, the truth of a counterfactual conditional with proposition P1 as the antecedent and with proposition P2 as the consequent can be trivially derived from the truth of P1 and P2. Since the analysans of (SCA) is a counterfactual conditional, it follows that, for any two true propositions, there is a true counterfactual conditional linking the two, and as a result that there is a true disposition ascription (as disposition ascriptions are, on (SCA), equivalent to counterfactual conditionals). The problem is that the two propositions that we combine in order to form a counterfactual conditional can be totally unrelated to each other. Take the true propositions ‘I am typing on a keyboard’ and ‘I cross my legs’. On the standard semantics for counterfactual conditionals, from the truth of these two propositions follows the following true proposition:

(a) If I were to be typing on a keyboard, I would cross my legs.

If (SCA) is true, (a) would in turn be equivalent to the following proposition:

(a*) I am disposed to cross my legs when I am typing on a keyboard.

It doesn’t take much reflection to realise that this is not necessarily right, as it could be a pure coincidence that I perform these two actions in succession.
Suppose now that it is indeed not a habit to cross my legs every time I am typing on a keyboard, as there are moments when I am typing on a keyboard but do not cross my legs, two other propositions thus being true of me at the relevant times: ‘I am typing on a keyboard’ and ‘I do not cross my legs’. The following counterfactual thereby follows:

(b) If I were to be typing on a keyboard, I would not cross my legs.

If (SCA) is true, then, in addition to being disposed to cross my legs when I am typing on a keyboard, I would also be disposed not to cross my legs when I am typing.

One could reply that having two contrary dispositions could happen, and, as we will later see, I would like to accept the possibility. Yet, we could imagine cases with two propositions that are even more unrelated. Replace the proposition ‘I am typing on a keyboard’ with the proposition ‘Madonna sings ‘Like a virgin’’. We now get the following disposition:

I am disposed to cross my legs (or not to cross my legs) when Madonna sings ‘Like a virgin’.

Needless to say, this is an unacceptable consequence of (SCA) if one accepts the relevant semantics of counterfactuals.

To conclude, even if (SCA) could capture some disposition ascriptions—the canonical formulation given above could well be adequately analysed by (SCA)—it fails to provide an adequate analysis of everyday disposition ascriptions, namely the ones that we are interested in.
I do not intend to propose my own analysis of disposition ascriptions, as I am sceptical of the possibility of giving one that adequately captures all cases (Section 10). I think this is not needed, however, as the search for an adequate analysis is sufficient to give us important insights about the nature of dispositions. In the next two sections, I will discuss further issues that could reinforce our scepticism.

8. Are dispositions an intrinsic matter?

We have seen that there is a temptation to think that dispositions are an ‘intrinsic matter’ (Lewis, 1997). However, it is questionable whether they are necessarily intrinsic. Are all dispositions intrinsic to their bearers? Although some would like to preserve the intrinsicness of all dispositions (e.g., Lewis, 1997, Molnar, 2003), there is a growing consensus that extrinsic dispositions are possible (e.g., McKitrick, 2003, Bird, 2007, Manley & Wasserman, 2008). According to Jennifer McKitrick, a disposition is intrinsic, roughly, “if anything that has it has it regardless of what is going on outside of itself.” (McKitrick, 2003, 158) Perfect duplicates must share them, as perfect duplicates must share all of their intrinsic properties. A disposition is extrinsic, by contrast, when the having of it depends on factors outside of its bearer. As a result, perfect duplicates can differ in their extrinsic dispositions. Another way to cash out the distinction is by saying that, while intrinsic dispositions’ instantiation does not depend on factors outside of their bearers, extrinsic dispositions’ instantiation does (Hoffmann-Kolss, 2010).

The examples that McKitrick gives of extrinsic dispositions are my key’s disposition to open my front door, Bill Clinton’s disposition to be recognized, and vulnerability or the disposition to get hurt. The question is whether such cases constitute genuine cases of extrinsic dispositions. If the answer is yes, then Lewis’ appeal to the intrinsicness of dispositions in order to counter finkish cases fails, and (RCA) is false. And if it is possible to establish the existence of
extrinsic dispositions, is there a good argument to the effect that all dispositions are actually extrinsic? We have seen that (SCA) requires an affirmative answer to this question, as it rules that all dispositions are extrinsic.

There are at least three possible views regarding the existence of intrinsic and extrinsic dispositions:

*Extrinsic disposition monism:* All dispositions are extrinsic and something like (SCA) may be true.

*Intrinsic disposition monism:* All dispositions are intrinsic and putative extrinsic dispositions do not actually exist.

*Disposition dualism:* Some dispositions are intrinsic and some dispositions are extrinsic.

I will consider each of these views in turn.

*Extrinsic disposition monism*

If all dispositions are extrinsic, then it seems that all the counterexamples to (SCA) can be blocked. An objector to (SCA) could ask us to consider the case of a fragile vase that would lose its fragility if it were dropped. Assuming here that *being fragile is being disposed to break when dropped*, the glass can nonetheless be said to be fragile at a time when it is not dropped, even though it wouldn’t break if it were dropped. Now, the friend of (SCA) could reply that, since all dispositions are extrinsic to their bearer, a better way to describe the case is by saying that, since the glass wouldn’t break when dropped, it is actually not fragile at any moment it is not dropped.

The problem with this response is that the friend of (SCA) is simply begging the question against the objector, by assuming *extrinsic disposition monism*. It is a
consequence of her view, and so cannot be a justification for it. Taken in itself, however, the view is implausible, precisely because it is very odd to re-describe situations such as the one above in the way the friend of (SCA) does. In those cases, the counterexamples are effective because it is highly plausible that the dispositions we are considering are intrinsic. Actually, a good explanation of the intuitive force of the counterexamples is that the dispositions in question are intrinsic, and a good test for it is to ask whether a perfect duplicate of the objects in question has the disposition. In the case above, it seems that a non-finkish perfect duplicate of the vase would definitely be fragile. As a result, finkish cases do constitute a counterexample to (SCA). 38

Intrinsic disposition monism

This response is defended, among others, by George Molnar (2003). According to Molnar, there are no extrinsic dispositional properties, only extrinsic dispositional predicates (Molnar, 2003, 108): “All dispositional and extrinsic predicates that apply to an object, do so by virtue of intrinsic powers borne by the object. All truths about the powers of objects have only intrinsic properties as truthmakers.” (Molnar, 2003, 109)

On such a view, I can speak truly when I ascribe to my key the disposition to open my front door. It is just that what makes true my statement are intrinsic dispositions of both my key and the lock of my front door, respectively. An account of vulnerability, to take another of McKitrick’s examples, could be that X is vulnerable just in case the complex situation of which X is a part has an intrinsic disposition to produce Y—the hurting of X, say. Extrinsic disposition ascriptions, on the view defended by Molnar, would be ways to focus attention to a part P of a broader context that we are particularly interested in, although it is not strictly speaking a disposition of P that does the work. They here primarily serve a pragmatic role.

38 Bird (2007, 31) agrees.
Another way of arguing for the *intrinsic disposition monist* view is by claiming that extrinsic dispositions are internal relations whose nature is entirely determined by the nature of the *relata* (like ‘being taller than’). A claim about two people’s joint disposition to lift a very heavy object, for instance, is true in virtue of the dispositions that each person intrinsically has. Similarly, the power that I have to lift a heavy object may depend on my arms’ respective powers. So the notion of extrinsic disposition may merely be practical and therefore may not require the addition of extrinsic dispositions in our ontology.

*Disposition dualism*

It is one thing to say that extrinsic dispositions, if they existed, would be ontologically dependent properties; it is quite another thing to say that they do not actually exist. McKitrick can readily accept a supervenience relation between extrinsic and intrinsic dispositions. Recall, moreover, that it is possible to claim that a given kind of entity exists while denying that it figures among the ultimate constituents of reality. Unless we accept the idea that all that exists are such ultimate, basic constituents, and that complex entities do not, strictly speaking, exist, we have so far no good reason to deny that extrinsic dispositions do not exist if we also accept that intrinsic ones do. An acceptance of extrinsic dispositions is compatible with the claim that intrinsic dispositions are more fundamental than extrinsic dispositions.

There is no harm, at least in our context, to accepting both intrinsic dispositions and extrinsic dispositions, as long as we are open to the thought that extrinsic dispositions may have intrinsic dispositions as constituents. This

---

39 I do not wish to suggest that someone dealing with the question of what are the ultimate constituents of reality will find extrinsic dispositions among them. In fact, as should be clear from the text, my task here is not to show that dispositions in general, intrinsic or extrinsic, are among such constituents.

40 One may wonder whether the view accepted here is genuinely dualist. Dualism about mental and physical properties claims that mental properties are entirely independent from physical properties. This means, among other things, that mental properties cannot be composed of physical properties (and *vice versa*). In our context, by contrast, it is accepted that extrinsic dispositions may be entirely composed of intrinsic dispositions. I take it that the view is dualist
however makes things more complicated if we want to give a general account of disposition ascriptions.

9. Intrinsic finks, masks and mimics

I would like to mention a further difficulty for the analysis of disposition ascriptions. Recall Lewis’ strategy to appeal to the intrinsicness of dispositions in order to accommodate cases of finks. If an object has a disposition whose manifestation is finkish, we can explain why the object really does have the disposition by saying that a perfect duplicate of it would have it too. We would then be able to say that the interfering factors, being extrinsic to the object, are irrelevant to whether the object has an intrinsic disposition. The difficulty arises when we contemplate the possibility of intrinsic finks, masks and mimics. I will focus on intrinsic masks, but I’m confident that cases of intrinsic finks and mimics can easily be found (Clarke, 2008, 2010, Everett, 2009).

Intrinsic masks, Molnar says, “are everywhere once you start looking at them.” (Molnar, 1999, 5) “The Greeks seem to have had a fascination for such cases: Tantalus, whose ability to drink was masked by his disposition to cause all fluids he approached to evaporate; King Midas, whose disposition to turn everything he touched into gold unfortunately masked his ability to nourish himself.” (ibid.) Other examples of intrinsic masks are not difficult to find:

Knowing French: Assuming that knowing French is an intrinsic dispositional property and that it is to be defined as the disposition to correctly answer questions asked in French, I cannot answer a question in French when I am asleep. Being asleep being plausibly an intrinsic condition, it would count as an intrinsic mask.

in the weaker sense that it accepts two kinds of dispositional properties when other views accept only one.
*Magic Wrap:* A fragile glass being enchanted by a sorcerer in such a way that, were the glass to be struck, the enchantment would cause bubble-wrap to appear around the glass (Ashwell, 2010, 9).

*Opposite dispositions:* I can like and dislike the very same thing. “Liking it, one has an attraction toward it, a disposition, perhaps, to seek it, acquire it, accept it, or consume it. Disliking it, one has a repulsion from it, a disposition, perhaps, to avoid it, reject it, etc. (…) When one does seek the thing in question, it isn’t one’s ambivalence that is manifested, it is one’s liking; and similarly (mutatis mutandis) with regard to when one avoids the thing.” (Clarke, 2010, 154) The manifestation of one disposition masks the manifestation of the other.

Arguably, a perfect duplicate of the objects in question would also contain the mask. A question is whether it is correct to describe such cases as involving a disposition that is masked by an intrinsic factor, as opposed to removed by it (and the corresponding disposition ascription falsified). The problem is to find a criterion that can distinguish between a case where a disposition is masked by an intrinsic factor and a case where the disposition is just plain absent.

Several writers are sceptical about the possibility of intrinsic masks, finks, and mimics (Choi, 2005, Cohen & Handfield, 2007). Here’s, for instance, what Handfield and Bird say: “We can conclude therefore that as regards both fink and antidote cases if S contains an intrinsic fink or antidote to some disposition D, then S does not possess D.” (Handfield and Bird, 2008, 291)

Although it may be true of certain types of dispositions, generalising this move is problematic, however. The first thing to say is that we want to preserve the possibility of extrinsic interfering factors and therefore the existence of intrinsic dispositions. A glass that is wrapped up in packing material can still be fragile;

---

41 Such authors mainly deny intrinsic finks, but I think their point applies to intrinsic masks as well.
the extrinsic interfering factor—the packing material—does not remove the fragility of the glass. In order to establish this point, it is customary to appeal to the point that (nomic) duplication should preserve the possession of dispositions when they are intrinsic (Choi, 2005). Since an intrinsic duplicate of the wrapped glass can obviously be fragile, the wrapped glass is fragile too.

The next thing to notice is that, if such a thought were accepted as a way to test whether a given object has a disposition, we would then be forced, it seems, to deny that an object’s disposition can be intrinsically masked. For a duplicate of it would also have the mask and therefore would not exercise the putative disposition; the test’s result is that the object does not have the disposition in question, at least not anymore. So, while an intrinsic duplicate test shows us that intrinsic dispositions can be extrinsically masked, it casts doubt on the possibility of intrinsic dispositions that are intrinsically masked.

There are reasons to resist this conclusion, however (Ashwell, 2010, Clarke, 2008, 2010). The first question is: why believe in the possibility of extrinsic masks? Take the case of the fragile glass wrapped up in some packing material. Why do we describe the packing material as preventing the fragility from manifesting itself as opposed to preventing the object from being fragile? It is because the packing material’s very function is to make sure that the glass, which is otherwise fragile, doesn’t break if struck. In other words, the best explanation of the role the packing material plays is that the glass is fragile. A competing explanation is that the glass would be fragile if it wasn’t wrapped up (Ashwell, 2010, 9). According to it, the role of the packing material is to prevent the glass from being fragile.

However, we have seen that, in the case in question, it is difficult to deny the fact that fragility is intrinsic. If the would-be explanation is true, fragility becomes an extrinsic matter, as two duplicates may differ in whether they are
fragile. Given that fragility is an intrinsic matter, the would-be explanation should be rejected as an explanation of the role of the extrinsic protector.

Now, we may consider this kind of explanation to deal with putative cases of intrinsic masks. Given that two intrinsic duplicates would not manifest the putative disposition, a correct description of the situation would be that the presence of the mask prevents the object from having the disposition. In the enchantment case above, for instance, one could argue that, since an intrinsic duplicate of the glass would not break, the glass cannot be fragile; the enchantment’s role is only to remove the fragility, or at any rate prevent it from being possessed by the glass. The problem, Ashwell argues, is that, if this explanation is to be applied to the enchantment case, there is reason to apply it to the normal—extrinsic—masking case as well. For both the enchantment and the packing material prevent the glass from breaking in exactly the same way: “the protective wrap plays the same role in both cases” (10). “If the protected but non-enchanted glass does not remain fragile when wrapped, then we lose a paradigmatic example of an extrinsic mask, and so it seems we should be suspicious as to whether masking of dispositions is possible at all.” (10). Ashwell concludes: “There is no principled way to exclude cases if intrinsic finking and masking without giving up on finks and masks altogether.” (17)

Another thought that vindicates the possibility of intrinsic masks, finks and mimics is that, whenever a disposition is intrinsic, it is in virtue of its bearer’s having a certain property, dispositional or otherwise, that is intrinsic. A question to ask whenever confronted with a putative case of interference is, does the object have the relevant property, or does it lack it (Clarke, 2008, 2010)? Such a ‘constitution test’, applied to normal cases of interference, seems to yield the right results. A live wire that is finked loses the property in virtue of which it is live, while a fragile glass that is wrapped up keeps the property in virtue of which it is fragile. Now, the test appears to allow cases of intrinsic interference. In the case of knowing French, for instance, we could ask whether the brain structure that is responsible for my knowing French is still present when I am
asleep; plausibly, the answer is affirmative. Moreover, in Ashwell’s enchantment case, it is not absurd that the glass keeps the property in virtue of which it is fragile, the same way it keeps that property when it is merely wrapped up in some packing material. I should emphasise that the constitution test doesn’t supersede the duplication test (Clarke, 2008). It is used where the intrinsic duplicate test fails to provide a clear answer to the question whether a given object possesses a disposition.

Given these arguments, and our commitments about dispositions, it seems that there is no good reason for us to deny the possibility of intrinsic finks, masks and mimics. It will later be clear that the acceptance of such cases can help explain certain psychological phenomena.

10. Towards pluralism

We have seen that, in many cases, (SCA) fails because the dispositions in question seem to be an ‘intrinsic matter’. However, we have also seen, with McKittrick, that if we are to accept dispositions as real properties of objects, we should accept both intrinsic and extrinsic dispositions. Accepting both intrinsic and extrinsic dispositions is, as a result, incompatible with a comprehensive analysis of disposition ascriptions appealing to the intrinsicness of dispositions. And given that (SCA) (i.e., a conditional analysis not appealing to any properties) was found mistaken, it seems that no unified analysis is in the offing.

My suggestion is to take a pluralist perspective on disposition ascriptions and acknowledge that there may be different, at least partial, analyses for different kinds of dispositions. At least three general kinds, given our discussion, can be posited:

(a) Some disposition ascriptions are analysable in terms of an intrinsic property doing some causal work (e.g., fragility). This property is objective and
can be possessed at times at which it is not manifested. Let’s call it simply *intrinsic disposition*.

(b) Some disposition ascriptions are analysable in terms of an extrinsic/relational property (e.g., vulnerability). This property is objective, can be possessed at times at which it is not manifested, and may supervene on intrinsic dispositions. Let’s call it *extrinsic disposition*.

(c) Some disposition ascriptions are analysable without mention of dispositional properties at all, and (SCA) may be true of them. Let’s call the relevant kind of disposition *Humean disposition*.

Given our foregoing discussion, there is no need to elucidate (a) and (b). (c), by contrast, requires a bit of explanation. What I have in mind here are primarily cases of disposition ascriptions made by committed Humeans, such as Gilbert Ryle, who do not think that there are dispositional properties at all, over and above the relevant conditionals, and who use dispositional predicates in a purely behaviouristic fashion. I suspect, however, that such a use can be found even in contexts where no commitment to (SCA) or anything like it is implicitly accepted by the relevant parties. Taking a case where a given disposition is said to be mimicked, it still makes sense, it seems, to say that the object is *disposed* to (or, better, *tends to*) act in a certain way if subjected to a given stimulus. Similarly, it seems to make sense, in agreement with Choi, to say that a glass that is wrapped up in some packing material is *not* disposed to break when struck. More dramatically, it makes sense to say that I am disposed to cross my legs when I am typing on a keyboard in order to capture a certain regularity or pattern in my behaviour. Ascriptions of Humean dispositions capture such patterns without making any claim about the facts in virtue of which such patterns obtain.

Such uses of disposition ascriptions, it seems, depend on our epistemic interests and capacities of observation, serving the purpose of making cognitive links between distinct events. Humean dispositions are thus not fully objective; there
is no such thing as a Humean disposition that is undetected. In contrast with intrinsic and extrinsic dispositions, furthermore, Humean dispositions cannot really be said to exist independently of their manifestations. At any rate, the thought strikes me as barely intelligible. Claiming that Humean dispositions exist is not implausible; what is implausible is to claim that all dispositions are Humean.

I would like to remain silent on whether (a) and (b) can be analysed in terms of conditionals. Given the foregoing discussion, I think scepticism is warranted. Furthermore, I do not wish to deny that ascriptions of dispositions can be analysed in a general way at all. What I am sceptical about is the prospect of an analysis that is both general and informative about the nature of dispositions. A more fruitful project is, I think, to focus on the various dispositional notions and ask what work they can do in our explanations (Bird, 1998). Some disposition ascriptions, it seems, are more explanatory than others. Quite plausibly, indeed, an appeal to an intrinsic disposition tells us more about the nature of a given object than a mere appeal to the patterns we see in its behaviour; it may be precisely because it has an intrinsic disposition that we can observe such patterns. As a result, some dispositions may ultimately be eliminable in our best account of reality. I am confident that Humean dispositions are; I suspect that at least some extrinsic dispositions are; and I am strongly confident that many intrinsic dispositions aren’t. I believe some of these ineliminable dispositions are psychological.

11. The case for intrinsic psychological dispositions

In this section, I will motivate realism about certain intrinsic psychological dispositions. In the next section, I will motivate realism about certain extrinsic psychological dispositions. Realism, let’s remember, is here the view that dispositional discourse cannot as a whole be paraphrased in the way the reductionist suggests (by analysing away dispositional properties) and that
least some dispositional properties exist (in the following, certain psychological dispositions).

According to Richard Wollheim (1999), mental dispositions have ‘psychological reality’ (4). What he means by this is that ascriptions of mental dispositions are not merely made true by the presence, actual or possible, of certain events. When I attribute to you the dispositional belief that the earth is spherical, I am not merely stating facts about how, among other things, you would answer the question ‘What is the shape of the earth?’ in certain situations. I am attributing to you the possession of an entity, a belief, that exists over and above these facts, and that stands in an explanatory relation to these. It is *because* you believe that the earth is spherical that you would answer correctly the question above, *not* because you would answer correctly the question that you can be said to believe that the earth is spherical. In short, believing is not a Humean disposition.

Another kind of psychological disposition that seems resistant to being eliminated in favour of other entities such as mental events is *desire*\(^4^2\). When I attribute to you the strong desire to be happy, I am not merely saying that, whenever you have the occasion to find happiness, or in a suitable proportion of such occasions, you would take your chances. I am saying something intuitively *deeper* than this. You desire to be happy even if, most of the time, you would not take your chances, even if, perhaps, you *never* take your chances. It may be because, being unhappy and hopeless, you fail to notice the relevant occasions. The presence of the desire explains why you would take your chances; it may also explain your misery when happiness is not around the corner (it is because your desire is unsatisfied that you are unhappy). Furthermore, it can explain why we can hope and be confident that our miserable friend will find happiness again; most people deeply desire to be happy, and as a result have the power to

\(^4^2\) At least desires of the dispositional sort. Naturally, I allow some desires—such as the occurrent desire for ice cream—to be mental events. The same may not be true of beliefs, however; beliefs, properly called, may never be conscious, as opposed to judgements, plausibly construed as the expression of beliefs and not beliefs proper (see Crane, 2001, 105-106).
take chances whenever the occasion presents itself, and that even if they are for the most part unhappy. Psychological dispositions such as beliefs and desires, it seems, are ineliminable parts of the explanation of our mental lives. This is not surprising, given the widespread acceptance that they should be fundamental to any ontology of the mind. Given the construal of realism given above, we should be realists about them.

To be fair, the foregoing discussion would probably not compel an error theorist about the mind (the so-called ‘eliminative materialist’) into accepting realism about psychological dispositions. The present point is more modest. It only suggests that beliefs and desires, due to their apparent explanatory value, have a privileged position in our common-sense thinking about the mind, and therefore that we should be realists about them if we are realists about the mind. If we are not error theorists (i.e., eliminative materialists) about mental states across the board, we shouldn’t be error theorists about beliefs and desires. In the following chapter, I will argue that the same sort of consideration applies to sentiments.

12. The case for extrinsic psychological dispositions

We have seen earlier that it is possible to ascribe to an object O a dispositional property whose existence depends in part on entities that are external to O. Although the disposition may be real in the same way typical intrinsic dispositions are, because (say) it is intrinsic to the complex pair of objects, it is not intrinsic to either constituent object; it supervenes on their respective properties as well as the properties of the situation. Some psychological dispositions may be like that. I call them extrinsic (or ‘shallow’) psychological dispositions.

An example may be a disposition to make correct calculation when aided by a calculator. The disposition is constituted by (i) a person, (ii) her ability to use

---

43 For a dispositional account of desire along realist lines, see Ashwell, 2009.
calculators, and (iii) a working calculator. The disposition is intrinsic to the entire set-up, but it is still a psychological disposition, as it is to the agent herself that we are likely to ascribe it (‘X is disposed-to-make-correct-calculations-when-using-a-working-calculator’) and it is in virtue of at least one psychological disposition that she possesses it (e.g., the disposition to properly make use of calculators). Nevertheless, the truth of the disposition ascription is not determined solely by facts about the agent—a working calculator must assist her—hence the corresponding disposition is extrinsic to her.

Such dispositions are psychologically shallow because they do not exist if some things extrinsic to their bearer do not exist; it is as part of a larger whole that the agent can be said to possess the disposition. And it is only when the property of being part of such a larger whole is instantiated that we can confidently say that the disposition in question exists. When not in such a situation, the most that can be said is that the agent has the disposition to contribute to the exercise of a disposition that an arrangement of objects including her would have if it were put together. This is not the same for intrinsic psychological dispositions; ‘X is disposed to make correct calculations when asked’ is true in virtue of something totally intrinsic to X—that she is good at maths, say—regardless of whether anybody ever asks her to count.

As I argued in Section 8, there is no harm in accepting the existence of extrinsic dispositions in addition to intrinsic ones. We have seen however that such dispositions may supervene on intrinsic dispositions. Does that mean that extrinsic dispositions are ultimately eliminable in our explanations? Perhaps. Perhaps indeed we shouldn’t be realists about them in a sense that is stronger than the one accepted here. Still, I think that speaking in terms of extrinsic dispositions, especially when it comes to psychological dispositions, is sometimes explanatorily useful, as it does not force us to say precisely what are the intrinsic properties other than the person’s that are doing the work and make us focus on what is important to us, namely the person herself.
13. Conclusion

The main task of this chapter has been to argue that the case for a reductive analysis of dispositional discourse that analyse away the apparent property-status of dispositions is at best only true of a minor subset of disposition ascriptions, and that many disposition ascriptions should be captured in realist terms. Disposition ascriptions, for the most part, purport to refer to genuine dispositional properties that are causally efficacious and thereby explanatorily valuable.

Although I have not attempted to conclusively show that error theory is false regarding all the sorts of dispositions that we ordinarily appeal to, I have provided some prima facie reasons for thinking that at least some psychological dispositions play a genuine explanatory role in our mental lives. If we are realists about the mind, I suggest, we should be realists about psychological dispositions. In the next chapter, I will argue that, if we should be realists about psychological dispositions, we should be realists about sentiments. The question will then be whether sentiments are intrinsic or extrinsic dispositions.\textsuperscript{44}

\textsuperscript{44} If one takes sentiments to be ‘Humean’ dispositions, then one doesn’t count as a realist at all, on my view. The realist, however, need not deny that some sentiments may be Humean.
Chapter Three

Emotional Dispositions

1. Introduction

From the previous chapter, we can conclude that the antirealist about emotional dispositions cannot argue for her view by appealing to the semantics of disposition ascriptions, as the latter appears, at least in part, to be based on realist assumptions; in other words, one cannot show that emotional dispositions are not genuine properties by defending a ‘global’ form of antirealism. Along with the antirealist’s ‘Humean’ dispositions, we have good reasons to think that there are dispositions that are objective, actual, and distinct from their manifestations, and that such dispositions come in two varieties— intrinsic and extrinsic. There is so far no conceptual reason to reject realism about emotional dispositions. An appeal to conceptual considerations only, however, cannot establish realism—after all, realism says that there are in the world such things as emotional dispositions, and that such things are causally efficacious, not just that if emotional dispositions existed, this would not be in tension with common-sense thinking. Recall from Chapter 2 my construal of realism. In order to justify our realism, we must not only show that the having of an emotional disposition depends on facts over and above the truth and falsehood of the conditionals that they are typically associated with, but also show that emotional dispositions can play an explanatory role that, in contrast with Humean and perhaps some extrinsic dispositions, is not eliminable. In short, we must show that error theory about emotional dispositions is false. In this chapter, I will consider, and show the merits of, the hypothesis that at least some emotional dispositions are intrinsic dispositions that have great explanatory value.
This further argument is needed, for it may turn out that emotional dispositions construed as intrinsic dispositions do not have as much explanatory value as emotional dispositions construed as Humean, or that the facts they allegedly explain are in no need of explanation, or that since they cannot exist over and above the mental events that they allegedly produce, we should at most construe them as Humean. Furthermore, if there are no good reasons to believe in emotional dispositions as intrinsic properties of persons, distinct from their manifestations, the error theorist could well suggest a revision of common-sense if it is realist about emotional dispositions—common-sense could surely be wrong on such a matter—towards the acceptance of a more attenuated form of emotional disposition, identified with (say) a certain kind of pattern of mental events. So, even if common-sense requires realism about emotional dispositions, it could revise its beliefs in a way that better matches the facts.

The goal of this chapter is to show that emotional dispositions in the realist sense can help us solve puzzles and account for some psychological phenomena in a better way than they would if construed in the reductionist sense. I will take explanations employing an antirealist theory of dispositions to appeal either to individual emotional episodes or to connections both between emotional episodes and between emotional episodes and other kinds of mental state, where such connections are causal or structural (i.e., pattern-like). The realist, by contrast, takes emotional dispositions to be intrinsic properties of persons, distinct from their manifestations but modally connected to them. Besides accounting for the relationship between emotional dispositions and occurrent mental states, we will see that the realist can answer a number of why-questions (i.e., she can provide explanations) in a more satisfactory way than the antirealist. In other words, the introduction of a new entity in the ontology of the theorist of emotion can help her answer a number of questions that would be difficult to answer otherwise.
2. Psychological and non-psychological explanations

We have seen in the previous chapter that neither reductionism nor error theory are highly plausible views about dispositions. Certain disposition ascriptions should be interpreted along realist lines and certain dispositional properties seem explanatory. Still, it is important not to overstate the case that has been made. So far, all that has been shown is that emotional dispositions could be genuine dispositional properties, and not merely Humean dispositions. However, one could disagree with such a construal. On the one hand, one could insist that talk of emotional dispositions in fact purport to refer to Humean dispositions only, namely collections of (possible and actual) mental events—one would here be arguing for a local form of reductionism about dispositions. On the other hand, one could claim that, although the relevant ascriptions purport to refer to genuine psychological dispositions, these entities do not in fact exist—thereby defending a local form of error theory. One reason why we may be tempted to deny the existence of emotional dispositions construed realistically is if we think that all the explanatory work such entities may do can be equally done by entities that we already accept, for instance the emotional episodes themselves. In this chapter, I will argue that good explanations of certain psychological phenomena are made possible once we accept emotional dispositions in our ontology of the mind.

Another reason why we may want to deny the existence of emotional dispositions is that, even assuming the existence of intrinsic dispositions, such things could never strictly speaking count as psychological. The antirealist about emotional dispositions could well accept that the events that occur in someone’s mind can be further explained by underlying dispositions of the subject; however, she may deny that that kind of explanation is genuinely psychological. On a certain conception of the mind, indeed, what I have called ‘psychological’ dispositions are best conceived as subpersonal entities that should not play a role in psychological explanations. So, according to the proponent of this view, if irritability, caring, and love are to be construed as intrinsic dispositions, then they are to be construed as non-psychological things,
which goes against the obvious thought that such things, if anything, are psychological. This, in turn, would motivate the thought that at best talk of emotional dispositions should be paraphrased into talk of collections of events. (We will see in the next chapter why one might think it important that entities such as caring and love count as psychological.)

In this chapter, I will assume that the kinds of explanation that realism about emotional dispositions allows are genuinely psychological, and will claim that they are better than those proposed by the reductionist. This is assuming that scepticism about psychological dispositions (*qua* psychological) is unwarranted, a claim for which I will provide some support in the next chapter.

3. Variety of emotional dispositions

In the previous chapters, I noted an ambiguity in the term ‘emotion’ between an occurrent reading and a dispositional reading. Examples of ‘emotions’ in the dispositional sense—what I call sentiments or emotional dispositions—include irritability, (long-term) anger, and (on some accounts) love (Goldie, 2010).

The list is so far restricted to entities that, though not occurrent, sometimes count as emotions. However, what I have in mind here is not restricted to such cases. On my view, an emotional disposition is a disposition to undergo a range of affective states such as emotions, moods and (perhaps) desires in various situations. This definition is general enough to count as emotional dispositions entities that we wouldn’t call ‘emotions’ in any sense of the term. Any dispositional state that is, at least in part, a disposition to undergo affective states is, or involves, an emotional disposition.

On such a very broad construal, emotional dispositions include entities such as one’s love for one’s partner, one’s value for justice, one’s concern for one’s parents, one’s compassion (the character trait) towards people in need, as well as one’s neurotic personality. For sake of simplicity, I will talk as if what
emotional dispositions trigger are emotional episodes of the garden-variety sort, namely fear, sadness, guilt, and so on. However, we must keep in mind that this does not exhaust the class of things that should count as manifestations of emotional dispositions; whatever the range of affect-laden entities the dispositions can produce, the points I will make with respect to emotional episodes will mostly be applicable to that range as a whole. Note that the reductionist may be fine with saying that emotional dispositions such as the ones just cited exist; she will only deny that they are ontologically distinct from the mental events and behavioural responses that they allegedly produce. On the reductionist proposal, emotional dispositions such as love, irritability and concerns exist by, so to speak, riding on the back of their so-called manifestations.

Believing in emotional dispositions does not imply believing in just about everything that could be described as, or as involving, dispositions to undergo a range of affective states. For instance, if a certain scepticism about character traits is warranted (e.g., Doris, 2002, Harman, 1999), we may have reasons to deny their existence, and as a result to deny the existence of entities that are or involve emotional dispositions. The same point is true of psychological dispositions in general; not all kinds of psychological dispositions, even intrinsic ones, are equally explanatory, and as a result may be reducible to other kinds of dispositions or even be eliminable. Again, character traits may be an example (see Wollheim, 2003, 24). Moreover, if one is an error theorist about a certain class of psychological dispositions, and if such dispositions are partly though not wholly composed of emotional dispositions, one need not be an error theorist about the emotional dispositions in question. Let’s imagine that a character trait is a combination of two types of psychological disposition, one emotional, the other behavioural. Now, suppose that psychological data show that most people do not act in a way that we would expect if they really had the character traits that we attribute to them. Given our definition of character traits, it seems that we have some reason to deny that they exist, or at any rate to think that they are less numerous than we thought. Does that mean that the emotional dispositions that character traits are supposed to be composed of either do not
exist or are less numerous than we thought? Not necessarily, for the only
evidence we have is behavioural evidence. One can have the emotional part of a
character trait without the behavioural part, and vice versa, and so we can be
error theorists about character traits without being error theorists about either of
their component entities—after all, being an error theorist about character traits
could be only being an error theorist about the actuality (if not the possibility\textsuperscript{45})
of a certain \textit{combination} of entities that we otherwise have independent reason
to believe exist. I will assess scepticism about character in Chapter 5. In the
following, I will argue that emotional dispositions \textit{as a class} are things we
should be realists about, while leaving open the question whether we have
reason to deny the existence of a subclass of them.

Now, a classification of emotional dispositions is possible, and can be made
along two main dimensions in a way that is, at least with suitable translation,
compatible with both realism and antirealism:

(1) \textit{Manifestation}: Emotional dispositions can be either ‘single-track’ or ‘multi-
track’; they can have single manifestations or multiple manifestations. In
addition, emotional dispositions can differ in the types of mental states that
constitute their manifestations.

(2) \textit{Stimulus conditions}: What typically activates a given disposition? Actions?
Thoughts? People? Objects? Is there anything that is common to all the things
that trigger the disposition? For instance, manifestations of love for my partner
can be triggered by a variety of states of affairs, objects and events—her
physical presence, her physical absence, the fact that she is ill, the view of a T-
shirt she used to wear, the thought that she might die, and so on—but it seems
that what is common to all these things is their close relationship with my
partner herself. Talk of \textit{the} object of my love is therefore warranted. Other types
of emotional dispositions, by contrast, do not seem to behave in this way with

\textsuperscript{45} It is indeed quite clear that most antirealists about character traits do not think that character
traits are not possible in a metaphysically robust sense (i.e., they do not deny that character traits
are metaphysically or logically possible). Rather, on purely empirical grounds, they only deny
that they are ever instantiated in the \textit{actual} world.
respect to their stimulus conditions. Temperaments, for instance, are dispositions to slip into moods in a variety of contexts that may or may not have anything in common, and often for no reason in particular; this may explain why moods, hence temperaments, are sometimes thought of as a- rational (Deonna & Teroni, 2012, Ch.9).

I think some kinds of emotional disposition are standard, widely accepted examples (take love, compassion, and irritability, for instance), and my concern is to defend a realist conception of them. However, it will be clear that my view is compatible with the existence of countless emotional dispositions for which we do not have a concept or name—e.g., innate dispositions that, because of some pervasive factor in our contemporary world, rarely or never manifest themselves. Both realists and reductionists can agree that such things as loving, irritability and caring are (or involve) dispositions (see next chapter for an account of caring as emotional disposition). They will however differ significantly on how they construe dispositions. For the realist, love is an intrinsic property affording possibilities; for the reductionist, it is at best a pattern of emotional episodes.

4. Wollheim’s defence of realism about psychological dispositions

To my knowledge, only one author has explicitly defended something like the view that there are such things as emotional dispositions (with which he ends up identifying what he thinks we typically refer to when we use the term ‘emotion’); it is Richard Wollheim (1999, 2003). He does so by first arguing for the existence of psychological dispositions construed realistically. His defence is however too underdeveloped to constitute the last word, so it will be important to further develop Wollheim’s original thoughts.

46 And the error theorist, seen above, who wishes to revise common-sense by adopting the reductionist’s metaphysical account of emotional dispositions.
Wollheim’s target is antirealism, construed as making both a semantic claim and an explanatory one:

The view I have in mind is that which equates the ascription of a mental disposition to a person with a general prediction about what that person will do, or would do, in a certain range of circumstances. The disposition itself is taken to be no more than a pattern of such doings. ‘Do’ here is an all-purpose word, which covers thinking, feeling, acting. (4)

Wollheim notes two difficulties with such a view. First, the view does not make room for the fact that we seem to experience our dispositions indirectly, through their manifestations. For instance, we often seem to characterise our dispositions, on the basis of their manifestations, as being weak or strong or in conflict with each other. In addition, we also often think of dispositions in different people being (say) of comparable strength or in conflict. If dispositions were mere patterns of events, it isn’t clear how they could be either weak, strong or in conflict with each other. Wollheim says:

For [antirealism] has to hold that, when an observer ascribes to someone a conflict of dispositions, this merely reflects an uncertainty on the part of the observer about which of two or more predictions to make. [The antirealist] finds himself in effect relocating the conflict from the mind of the person to whom it is attributed, which is its natural site, to the mind of another, who is a mere outsider. (5)

Although Wollheim’s argument may seem rather underdeveloped (what does in our experience make us think that the underlying dispositions are weak or strong?), I presume that his goal is to show that the antirealist view is revisionary of our ordinary speech and understanding of psychological
dispositions. The antirealist could however accept the revisionary consequences if it turns out that she is right about the explanatory idleness of emotional dispositions as construed realistically.

The second difficulty is that the antirealist view of psychological dispositions “can make no sense of the explanatory value that dispositions are ordinarily thought to have.” (5) Wollheim points out that we ordinarily explain other people’s behaviours and mental states by positing entities like beliefs and desires that are responsible for their occurrence. He notes that this doesn’t mean that the antirealist denies an explanatory role to psychological dispositions; “but he does deny that their explanatory value comes from their being causes.” (ibid.) He continues:

As mere patterns in what people do, dispositions cannot have causal power. In consequence, when we connect what someone does with a disposition of his, and claim that the latter can explain the former, what we do, according to Ryle, is to subsume one thing that the person does under a series of things that he is in the habit of doing, and this is explanatory only in the limited sense that it takes away the oddity that the single happening might otherwise have. (ibid.)

What the antirealist denies, according to Wollheim, is the ‘psychological reality’ of dispositions.

This second criticism is in agreement with what I argued at the end of the previous chapter—that psychological dispositions are explanatory in the way Wollheim suggests. However, it is only once we have managed to reject the reductionist’s very starting point, namely her claim about dispositional discourse, that we can point to the explanatory value that this discourse has in everyday psychological explanations. After all, if the reductionist is right about
the semantics of disposition ascriptions, she could well say that we are confused when we grant psychological dispositions such an explanatory role, and that if we still want to grant it, it is our position that is revisionary, by positing entities that common-sense is not really committed to.

The work of the previous chapter was therefore necessary to show that ordinary dispositional discourse is not as the reductionist may portray it and that the realist view that psychological dispositions exist over and above the patterns of events that they give rise to is not a revisionary position.

5. Specification of realism about emotional dispositions

In this section, I would like to say a bit more what it takes to be a realist about emotional dispositions or sentiments. The view involves three claims—the first is an existence claim, the second is a claim about the relationship between emotional dispositions and outside observers, and the third is a claim about the relationship between emotional dispositions and their so-called manifestations:

Realism about emotional dispositions: (i) There are emotional dispositions (existence-claim), i.e. dispositions to undergo certain emotions and other affective states, that are independent both (ii) of any observer’s capacity of detection (observer-independence claim) and (iii) of the events (e.g., emotions) that, together with other conditions, they cause (manifestation-independence claim). As a result, several dimensions of variation between observer, disposition and manifestation are possible:

(a) A can think that B has an emotional disposition D, but B doesn’t possess D (i.e., A is wrong). [Consequence of (ii)]

(b) A can possess an emotional disposition D, but D doesn’t cause the events that are characteristic of its manifestation. Alternatively, A can acquire D
at $t_1$, lose D at $t_2$, but D doesn’t manifest itself between $t_1$ and $t_2$. [Consequence of (iii)]

(c) A series of mental events characteristic of emotional disposition D occurs in A, but A doesn’t possess D.

The reductionist denies one or more of the claims (i)-(iii). We have seen that she is free to say that emotional dispositions exist, so to accept (i). However, she is unlikely to accept (iii), as she thinks that if emotional dispositions exist, they supervene on things other than themselves—mental events and behavioural responses and patterns thereof being the most obvious possibilities. The denial of (iii) is likely to be based on a denial of (ii), although it need not be. The denial of (iii) could be supported by an appeal to parsimony—only one kind of psychological entity, instead of two, is ultimately accepted—rather than on any special, potentially controversial epistemological theory. Accordingly, my target opponent will be the reductionist who accepts (i) but denies (iii). I believe that a good reason to accept realism rather than reject it is that a number of psychological puzzles and phenomena would be solved or explained better than if we rejected it (i.e., if we rejected any of its clauses).

Before moving on, note that the definition I gave is neutral on whether emotional dispositions are intrinsic or extrinsic. However, it will become clear that a condition for a proper account of certain phenomena involving emotional dispositions (e.g., masking cases) is to assume that the latter are intrinsic to their bearers. This claim is nonetheless compatible with there being emotional dispositions that are extrinsic to their bearers.

6. What is the relationship between emotional dispositions and emotional episodes?

The realist and reductionist disagree over the role the notion of disposition is supposed to play (and the realist and error theorist may disagree over the role it
should play) in our thinking about the world. For the realist about both emotional episodes and emotional dispositions, emotional dispositions explain why we undergo the emotional episodes from time to time, and they explain them by causing them. For the reductionist (and the revisionary error theorist\textsuperscript{47}), by contrast, it is because we undergo certain emotional episodes from time to time that it is correct to attribute to us an emotional disposition. The notion of an emotional disposition, for the reductionist, is not explanatory; it is classificatory. To quote Shoemaker again, an emotional disposition is “a kind of conceptual-linguistic structural framework we overlay upon emotional reactions to render them intelligible” (2003, 94).

The reductionist is going to say that emotional dispositions are dispositions to undergo a range of emotional episodes in certain contexts, where such dispositions are understood as nothing over and above these emotional episodes. In the particular context of emotion, one may say that to have an emotional disposition is to exhibit a pattern of emotional episodes over time.\textsuperscript{48} To attribute an emotional disposition is to say (or predict) how a person would (or will) feel in given situations.

How can the reductionist account for psychological explanations that appeal to emotional dispositions? Everyday psychological explanations, it seems, are filled with such explanations. Why is David angry now? Because he is irritable. But if being irritable is just undergoing anger from time to time (or a similar story), it seems that the only thing that this answer manages to do is to tell us that there is nothing exceptional about David being angry now, as he is angry in many situations. Put another way, David’s current anger falls under a true generalisation about him and that is supposed to explain his current anger.

\textsuperscript{47} In the following, when I speak of the ‘reductionist’, I will refer to both the reductionist as understood so far and the error theorist who advocates a revision of ordinary emotional dispositional discourse in a way that reductionism becomes true of it.\textsuperscript{48} Helm (2001). See Chapter 1.
Still, the reductionist would have to deny that there is a deeper psychological explanation of why David is angry in many situations, why he has the relevant Humean disposition. It could be a sheer fluke, for instance, or there might be in these situations one particular stimulus that is present and that triggers the emotion. She could also appeal to the fact that these situations are approached with a sufficiently similar psychological context. For instance, she could say that David becomes angry in all the relevant situations because he believes that the world is against him and desires to be left alone, or because the people triggering his anger remind him of his parents, or whatever. Whatever the explanation, the reductionist claims that it need not appeal to any distinct emotional disposition in order to constitute a complete psychological explanation.

The realist, by contrast, thinks that the relevant psychological explanations can be incomplete. She claims that an appeal to similar situations and psychological contexts not including distinct emotional dispositions is sometimes insufficient to account for certain facts involving emotional episodes. What she claims, for example, is that there is something about David that can explain why he is angry in many situations, i.e. why a generalisation is true of him, namely that there is something in him, distinct from the anger that he feels from time to time, that causes it; this thing is his irritability. I will argue that construing emotional dispositions as intrinsic properties of persons, distinct from their manifestations, can help us explain a variety of psychological phenomena that occur in the emotional domain in a more satisfying way than done by the reductionist. My goal is indeed to show that there is a need for emotional dispositions in addition to emotional occurrences, and that emotional dispositions are not reducible to the occurrence, possible or actual, of emotional episodes in certain situations; they are as actual and real as the emotions that they give rise to.
7. Explaining mental events

In the introduction of this thesis, I mentioned the fact that some of us sometimes feel a need to explain the occurrence of the events we observe. Regularities between events are one case for which we would like an explanation. And we have seen philosophers who are not afraid of appealing to (presumably) unobservable entities such as laws of nature, higher-order universals, or causal powers in order to explain some or all of these regularities. Regularities are not the only case where the need for an explanation can be felt. Most, perhaps all, past events that we can think of are things that could have failed to happen. The fact that they occurred is a contingent fact about the world. As such, they seem to call for an explanation. Why did they happen when they could have easily failed to happen? Why did the car crash when the driver could have avoided the other car by directing the car slightly to the left? The fact that some kind of event failed to occur is also something for which we can need an explanation, especially when that event was expected. Why did the patient survive the disease when it was almost certain she would not?

Of course, all these events could be explained in a variety of ways. They could be explained in physical terms, chemical terms, biological terms, psychological terms, and so on. In this thesis, my concern is on psychological explanation. I think that, as events in general are things for which we can be in need of an explanation, mental events are things for which we can be in need of an explanation. Why did Sarah feel this way when the doctor told her that her father was ill? Why (or perhaps how) did Paul come to the conclusion that the earth is flat? And, of course, an explanation can be asked for mental events that failed to happen but could or should have happened. Why did Tony fail to feel sad upon hearing bad news about his wife?

Emotions, I have assumed, are mental events. Their occurrence is therefore as much in need of explanation as any other kind of mental event. Several explanations can be asked about the occurrence (or non-occurrence) of
emotions. Here are a few questions in response to which such explanations may be given.49

1) Why did some kind of emotion occur rather than another or none? For instance, why did David get angry with, rather than pleased by, the waiter?
2) Why does some psychological state of affairs about a subject’s emotions exist? For instance, why does David often get angry when witnessing an act of injustice?
3) Why does a state of affairs about a subject’s emotions fail to obtain? For instance, why does David never feel happy when meeting Andy?
4) Why does some state of affairs about one’s emotions still obtain? For instance, why is David filled with joy every time he sees his wife after some time away?
5) Why did some kind of emotion fail to occur (when it could have, or was expected to)? For instance, why did David fail to feel compassion at the view of a person in need?
6) Why did some kind of emotion fail to occur in situation of type S when this kind of event has typically occurred in situations of type S in the past? For instance, why was David not happy to see his wife when he came back from his last trip?
7) Why did some kind of emotion occur in situation of type S when it rarely occurred in situations of type S in the past? For instance, why did David feel joy instead of irritation (his usual response) in listening to a piece of classical music?

In the following, my aim is to show that each of these questions can be answered by appealing to distinct emotional dispositions and that the relevant explanations may be more complete than explanations not involving such dispositions. The pattern-based account of emotional dispositions (Chapter 1), not involving distinct emotional dispositions, will therefore be shown to be less satisfactory.

49 Most of the following questions are drawn from Steward (1997, 177-178) who, in the relevant pages, is concerned with the explanation of events in general.
Two main kinds of explanations of mental events are possible on the pattern-based account, depending on the theorist’s broader views about the mind. First, the pattern-based account of emotional dispositions may be special case of a pattern-based account of mental states in general. On such a view, all mental phenomena, including emotions, intentions, thoughts, but also beliefs and desires, are mental events (or clusters of mental events) organised in patterns (Nelkin, 1994). No underlying dispositional entities, except perhaps non-psychological ones (such as dispositions of the brain), can be appealed to in order to explain the occurrence or non-occurrence of such mental events, at least when psychological explanations are the aim.

Second, the pattern-based account of emotional dispositions may be combined with a general view of the mind that includes both mental events and mental dispositions. For instance, one could reduce sentiments while accepting that beliefs and desires (say) are genuine psychological dispositions. In such a case, one could explain the occurrence (or non-occurrence) of mental events by appealing to both other mental events and mental dispositions, as long as such dispositions don’t include sentiments.50

In the following, I’ll ignore the difference between both views and show that the realist about sentiments can provide explanations that are more complete than the ones given by the reductionist. The challenge for the realist about emotional dispositions is to show that there are cases where she can answer each of the above questions in a more satisfactory way than the reductionist, of either stripes, and no case where her answers are less satisfactory (for, after all, her ontology includes all the entities the reductionist accepts). Let’s now turn to

---

50 This is assuming that sentiments are not a species of belief or desire, in which case the difference between the realist and the pattern theorist would turn out to be, not a difference in ontology (the kinds of entities one accepts), but a difference in the kind of thing one takes sentiments to be (‘sentiments are patterns of emotions’ vs. ‘sentiments are dispositional beliefs/desires’). For now, I will mostly assume in the following that sentiments are neither beliefs nor desires. I think this assumption is innocuous given that, if they were, they would be of a special sort (see next chapter). In addition, it is implausible that all sentiments are beliefs or desires; irritability, for instance, is surely not a propositional attitude, and is at best belief-like or desire-like. In later chapters, I will tentatively tackle the question whether central cases of sentiments (caring and love) are belief-like or desire-like.
these questions, one by one. The upshot will be that, unless there are independent reasons to adopt an error theory about sentiments—perhaps based on a suspicion about psychological dispositions in general—sentiments seem to make a significant contribution to various psychological explanations.

8. Explaining emotional episodes

1) Why did some kind of emotion occur rather than another or none?
Consider the following simple case. Jane felt fear at the sight of an animal. Why did she feel this way? Relevant factors include the presence of the animal, facts about Jane’s ability to perceive it, and the presence of other psychological states. The question is whether it is plausible to think that Jane’s psychological states include emotional dispositions. Presumably, Jane believes that the animal exists. She also may desire to flee the situation. Perhaps she will entertain thoughts about the possibility of a tragic end and will wonder how her family will react to the news, and may judge or believe that continuing on living is a good thing and intend to fight the animal if needed.

Among such psychological factors, the most relevant to an explanation of Jane’s fear seem to be her belief in the existence of the animal, her thoughts about the possibility of a tragic end (if occurring before the emotion) and her judgement or belief that life is a good thing. Is there any work left to sentiments? If caring is indeed a species of sentiment (Ch.1 and Ch.4), it seems that the fact that Jane cares about her life should play a role in explaining her reaction as well. As we have seen in Chapter 1, the fact that someone judges that something is good does not seem to entail that she cares about it, and one certainly can care about something without judging that something to be good, or without judging anything for that matter (as, presumably, in human infants). It is therefore possible for Jane to judge that continuing on living would be a good thing yet to fail to care about her life. And if told that this is actually the case, we might find her reaction to the animal puzzling. Why did Jane fear the animal if she doesn’t care about her life anymore? Of course, her fear is not that puzzling; it would be
natural to assume that Jane still cares a little about her life, or at least about the absence of pain. But the fact that such a question makes sense suggests that caring for one’s life typically figures in the explanation of fear experiences.

It may be objected that the *desire to live* could explain the fear as well, and that an appeal to caring would be unnecessary here. The fact that Jane desires to continue on living surely explains her reaction. Does this show that the pattern theorist can appeal to such a desire to complete her explanation? It depends on what conceptions of the relevant desire are open to her. For instance, she could construe the desire to live as some fleeting mental event popping into Jane’s head. She could also construe it as a deeper disposition of hers. Whatever kind of thing the reductionist takes Jane’s desire to live to be, she cannot construe it as a species of *caring*, since she defines caring in terms of patterns of emotional reactions.\(^5\) As a result, she is unable to provide a causal explanation of Jane’s fear by appealing to the fact that she cares about her life. The realist, by contrast, can allow that the desire to live sometimes counts as a care, as long as it satisfies certain conditions (see next chapter for such conditions). In fact, so far her view is compatible with all sentiments being desire (or desire-like) states. Unless Jane’s desire to live is construed as on a par with ‘shallow’ desires such as the desire for ice cream or the temporary desire to quit philosophy to become a rock star after having gone to a concert, there is no reason to deny that an appeal to it could be equivalent to an appeal to the fact that she cares about her life.

The explanation that a reductionist may give for individual emotional episodes may therefore look less complete than the explanation given by the realist. But can she appeal to the fact that the fear is part of a pattern of emotions of a certain sort? Of course she can, but we might be convinced by Wollheim’s remark to the effect that such an appeal would only show that there is nothing exceptional with Jane’s emotion and would thereby fail to contribute to its explanation. In other words, an appeal to cares, whether desire-like or not, is

\(^5\) She may be able to construe the desire as part of the pattern constituting caring (Helm, 2001). This however would only show that the desire to live can be *part* of an instance of caring.
informative in a way that an appeal to patterns is not. Wollheim thus provides us with a prima facie reason to think that an appeal to sentiments can provide more satisfactory stories than the ones provided by the pattern theorist.

2) Why does some state of affairs about a subject’s emotions obtain?
Consider the case of Robert, who happens to feel many positive emotions towards his family over time. For instance, he feels happy when his children visit him, and misses them terribly when they are away. He is also very pleased and proud when his wife is successful at work. He would be devastated if anything bad were to happen to any of these people. One might wonder why Robert happens to feel in all these ways over time, why he instantiates the relevant pattern.

The realist about sentiments has a simple explanation: it is because Robert loves his family that he feels all the relevant emotions over time. The reductionist, by contrast, is forced to say either that mental states other than emotions explain the pattern or that the pattern has no deeper psychological explanation. If she were to choose the former option, she would have to provide us with details about how a certain collection of mental states can give rise to patterns of emotions that intuitively count as (e.g.) instances of love. She cannot, for instance, appeal to desires or intentions of the subject to feel the relevant emotions (Chapter 1). In such a case, she would at best have an explanation of the instantiation of a pattern resembling love but not of a pattern constitutive of love. I will not assess each possibility (do Robert’s beliefs and desires suffice to explain his love?), noting the simplicity with which an appeal to sentiments can explain to presence of many patterns of emotions (especially the ones that intuitively count as expressions of love, caring, and so on).

The second option—that of claiming that patterns of emotional episodes do not require any special sort of explanation—is problematic as well. The main problem with this strategy is that the pattern theorist is pressed to account for the difference between a pattern of emotional episodes that count as a sentiment (e.g., love) and an exactly similar pattern that merely mimics it (Chapter 1).
Another problem is that it is becoming quite hard to understand why the pattern theorist would insist on denying the reality of sentiments if she is willing to accept psychological dispositions in her ontology. It seems that, if one is willing to accept psychological dispositions in general, one has no reason to deny emotional dispositions. The main motivation for the pattern theorist to deny emotional dispositions is therefore to deny psychological dispositions in general. In the next chapter, I will provide reasons to reject a picture of the mind as consisting only of mental events (and collections thereof).

3) Why does a state of affairs about a subject’s emotions fail to obtain?

Why do I fail to instantiate a pattern of emotional episodes towards my neighbour of the sort Robert instantiates towards his family? Because I don’t particularly love my neighbour. I don’t seem to think that he is unworthy of being loved or that he couldn’t become my friend. I don’t have any particular desires that would explain my lack of love. I just happen to fail to love him and that is why I don’t feel all the relevant emotions. This is one sort of explanation the realist can give for the absence of patterns of emotions.

Another sort of explanation the realist can give for absent patterns appeals to masks. Recall from Chapter 1 the possibility of sentiments, such as caring and love, without patterns. We saw the case of a father who, because of a depression, failed to feel the relevant emotions towards his family. We concluded that he may still love his family even during his depression. The realist about sentiments can now say that this is because sentiments are dispositions that can in principle be masked.  

The advantage of the realist view, therefore, is to allow for explanations of our emotional episodes in a way that does justice to our pretheoretical thinking on sentiments.

---

52 An interesting case would be where the entity that plays the masking role is another intrinsic disposition of the subject. See Chapter 2 for the possibility of intrinsic masks.
4) Why did some kind of emotion fail to occur (when it could have, or was expected to)?

Adequate explanations can be given, from a realist perspective, for milder cases of failures to feel. Take the case of Robert again who, one evening, fails to be pleased to see his wife. One explanation is that he stopped loving her at that very moment. This is an explanation that both the realist and the reductionist would be able to give. This explanation, however, would probably be wrong. The more likely explanation is that Robert failed to be pleased because of fatigue, lack of attention, stress, a bad mood, or any other distracting or obscuring factors. One might wonder precisely what role such factors play here.

According to the reductionist, all that could be said is that what they do is take the place of the emotional episode whose occurrence we were expecting. It could still be asked however why such a replacement is possible at all and why it does not falsify our ascription of disposition. The realist, by contrast, can easily appeal to the presence of masks and to the fact that moods, inattentiveness and tiredness are intuitively the kinds of things to play such a masking role (hence calling them ‘obscuring’ factors). Furthermore, if such masks are intrinsic to their bearers, as moods seem to be, it is safe to say that the dispositions that are masked are also intrinsic to their bearers.\(^{53}\)

At this stage, the reductionist could stress that emotional disposition ascriptions are not exceptionless generalities so that the ascription is not falsified if a few exceptions are present. If, by contrast, the inattentiveness (say) is so present that the relevant generalisation is not true of the subject anymore, we have seen that the reductionist ought to give up her attribution while the realist can retain it. Indeed, mere inattentiveness, even repeated, does not seem to be the kind of thing that by itself should make us deny the presence of a disposition. It is just that the person has a disposition that is more often than not masked by the inattentiveness, even though she doesn’t have the relevant Humean disposition (i.e., she is not disposed to feel a certain way in certain situations, in the

\(^{53}\) An extrinsic disposition that is intrinsically masked is impossible simply because an extrinsic disposition needs to be part of a larger whole in order to exist at all; at best what is masked is the disposition to become part of a larger whole.
reductionist’s sense). By allowing such psychological explanations, the realist therefore seems to have the upper hand over the reductionist.

5) Why does some state of affairs about one’s emotions still obtain?
Robert instantiates the same pattern of emotions towards his family as he did ten years ago. The realist is going to say that this is because he still loves them. The reductionist cannot appeal to Robert’s love because it would be equivalent, on her account, to appealing to the pattern itself; it would amount to saying that Robert still instantiates a pattern (i.e., still loves) because a pattern (an instance of love) is still present. She will therefore have to identify mental states, such as beliefs and desires, that would explain why the pattern hasn’t extinguished. Of course, she will find many that will partly explain its persistence, but not all of them will make it apparent that Robert still loves his family. For instance, it could be the case that it is because Robert desires to continue feeling the way he used to that a similar pattern is thereby present. But this can hardly justify our attribution of love to Robert on the basis of the relevant pattern. Of course, the reductionist could appeal to other desires of Robert’s in order to explain the relevant pattern in a way that makes it apparent that it is a pattern constitutive of love. Robert plausibly desires that his family flourish, that they do not live in poverty, that they live a long life, and so on and so forth, and these desires seem relevant to the explanation of the pattern, or at any rate to the occurrence of some of its members (e.g., the occurrence of a fear experience when Robert’s daughter hasn’t come back home yet).

The realist need not disagree here. Such desires make a contribution to the occurrence of the pattern. The problem is that the reductionist is forced to explain why these desires produce a pattern of emotions constitutive of love whereas desires such as the desire to have loving emotions don’t. Both classes of desires can in principle produce the relevant patterns of emotions, so both classes, on the reductionist’s account, should in principle be able to produce instances of love. The problem for the reductionist is that her account prevents her from appealing to the particular nature (including the content) of the

54 At least the one who thinks that the persistence of the pattern is in need of explanation.
relevant desires in order to explain the difference between a pattern of emotions that is constitutive of love and a pattern of emotions that merely mimics love, for on her account no such distinction is available. In contrast, the realist certainly can appeal to the relevant desires and claim that, as dispositional entities, they may be constitutive of love, and that it is because of their being constitutive of love that the patterns that they produce are expressive of genuine love (as opposed to merely mimicking love).

6) Why did some kind of emotion fail to occur in situation of type S when this kind of event has typically occurred in situations of type S in the past?

As with question 4 above, distinctive realist explanations can be given in response to this question. For instance, an underlying emotional disposition may have stopped existing, or it may be masked by factors of the environment, or it may be masked by the manifestation of another emotional disposition, and so on and so forth. For instance, Bob had always reacted with anger when someone speaks to him in a way that his father used to until he one day failed to be angry. This unexpected lack of reaction can be explained in a variety of ways. Perhaps Bob has come to terms with his past and from now on won’t be angry with people reminding him of his father. This hypothesis makes the prediction that Bob will continue to fail to feel anger in the relevant kind of situation. Another hypothesis is that Bob fails to be angry because he is in such a good mood that not even offensive remarks can reach him. In such a case, his disposition to get angry in the relevant situation is plausibly still there, albeit in a dormant form. Finally, Bob could have failed to be angry because the person in front of him is so friendly and lovable that the fact that he reminds him of his father is of very low significance to him.

7) Why did some kind of emotion occur in situation of type S when it rarely occurred in situations of type S in the past?

Finally, the realist about emotional dispositions, with her more inclusive ontology, can provide explanations that the reductionist would not be able to provide for cases of unexpected emotions. For instance, she can explain the occurrence of an emotion by appealing to the acquisition of a new disposition or
the manifestation of an already existing disposition that is otherwise masked most of the time.

Overall, the realist about sentiments seems to have explanatory resources that the reductionist lacks, and her view should be preferred accordingly. In the next section, I’ll show that the realist can accommodate a puzzling fact about love that the pattern-based account is sometimes praised for capturing, closing our defence of realism.

9. Multiple expressions of same-type emotions and same expression of other-type emotions

There is a phenomenon, hardly articulable in a precise manner, that nonetheless seems to be part of our ordinary thinking about emotion. It is the idea that two tokens of the same emotion type can intuitively ‘express’ something different in different contexts. Of course, two tokens of the same type can be about two different things; they can differ in the kind of object to which they are directed, although, as we saw in Chapter 1, they must have the same formal object. For instance, one’s anger in context C may be directed at a restaurant waiter whereas one’s anger in context C* may be directed at an instance of injustice. Here, the appeal to different (material) objects seems to provide an adequate explanation. However, these are not the cases that I have in mind here. Compare the case of feeling amusement at another’s embarrassment when that person is one’s lover with feeling amusement at another’s embarrassment when that person is unrelated to oneself. It seems that in the first case the amusement can be an expression of something like love while in the second case the amusement may be an expression of cruelty or schadenfraude.55 Positing a different object seems inadequate, or at any rate incomplete. Over and above what an emotion represents about the world, there is, to put it rather metaphorically, what the emotion says about the person undergoing it.

55 See Baier (1991) for a description of this phenomenon.
The realist about emotional dispositions seems to have an account of the thought at play here. Although the object of one’s amusement may be the same in both cases, its dispositional origin is different. The reductionist, by contrast, denying the existence of distinct dispositions, would have to appeal to the fact that the amusement is part of a certain pattern in one case and part of another pattern in the other case, given that all there are to dispositions are patterns. So, on this view, love would be a pattern distinct in kind from traits like cruelty (Helm, 2001, and Ch.1).

A different but related issue is how two emotion tokens of two different types can intuitively ‘say’ the same thing about the person undergoing them. For instance, it seems that both anger and sadness can constitute an expression of love, depending on the situation. The realist cashes out this claim by saying that the two emotional episodes both express love because they originate in the same disposition. No need to devise any complex story as to how seemingly disparate emotional episodes relate to each other. The view accounts for the fact that seemingly disparate emotional episodes are to be somehow ‘tied’ together in psychological explanations; the ‘tying’ relation in question is grounded in the fact that the relevant emotional episodes are produced by the same disposition (i.e., the disposition is some sort of ‘tokens-binder’\textsuperscript{56}). The reductionist, by contrast, would cash out this claim by saying that the anger and the sadness are both part of a pattern characteristic of love.

So it must be acknowledged that realism and reductionism (i.e., the pattern-based account) provide equally plausible accounts of a certain class of phenomena. Unfortunately for reductionism, however, realism has been found to be more plausible on other grounds, both in providing better psychological explanations and in avoiding the problem cases given in Chapter 1. The point of this section has been to show that the reductionist’s ability to account for the relevant class of phenomena cannot give us a reason to favour her view.

\textsuperscript{56} Thanks to Phil Letts for suggesting this term to me.
10. Concluding remarks

Why a realist defence of emotional dispositions (and psychological dispositions in general) when common-sense already takes them to be real things? First, because some people, including some working on emotion, may not be realists about dispositions in general. Second, because there are certain claims, descriptions, moves and arguments that only a realist can provide and that are very controversial if not supported by a realist theory. For instance, as we will see in Chapter 5, appealing to the presence of masks in order to save certain phenomena (e.g., the existence of character traits) in the face of scepticism can only be done if we accept intrinsic dispositions in our ontology. Alternatively, the appeal to the possibility of mimics can explain the possibility of emotional episodes that are not part of a pattern, contra some writers who may think that an emotion must be part of a pattern in order to be either an emotion at all or an emotion of a certain kind.57

My primary goal in this chapter has been to both motivate and provide an initial contribution to a certain research programme, namely one in which we investigate the place and role of emotional dispositions or sentiments in our emotional lives in particular, and in our lives in general. If what I have to say here is on the right tracks, sentiments deserve a more central place in an inquiry about the mind than they have been granted so far.

What I’ve done in this chapter is show that emotional dispositions construed as genuine properties can do important explanatory work, in contrast with reductionism and at least one form of error theory (an error theory about dispositional discourse as a whole). However, I have not shown that a local

---

57 This view may be therefore at odds with holistic accounts of emotion according to which in order to undergo an emotion at a given time, one must undergo other emotions at other times. The ‘must’ here can have a normative reading or a descriptive reading, and this is only against the latter that the present point is directed. Helm (2001) presumably has a normative reading, although he seems to slip into a more descriptive one when he says that “to fail to have these other emotions is for one’s response to fail to be a response to import—to fail to be an emotion at all”. (Helm, 2010, 251) My view is indeed compatible with the claim that being part of a pattern is a normative but non-essential property of some or all emotion types.
error theory about the relevant properties *qua psychological* is wrong. After all, we could have principled reasons to deny that dispositions construed realistically can count as genuinely psychological. This opens up the possibility of defending an antirealist account of emotional dispositions (and what I called ‘psychological’ dispositions in general) while accepting realism about dispositions in general. One could agree that emotional episodes typically can be explained by appealing to the presence of dispositional entities while denying that such entities are genuinely psychological. The same way one can be confident that mental events and processes are in fact produced by non-mental brain events and processes, one can be confident that such events and processes are caused by non-mental dispositional properties. In both cases, citing the brain or the underlying dispositional properties can show that mental entities do not float completely free from the rest of the world, but it does not clearly contribute to our business of providing plausible *psychological* explanations. Further argument is therefore needed in order to vindicate my account of sentiments.
Chapter Four
Caring, Attitudes, and Character

1. The realist-dispositional account of cares

We have seen in Chapter 1 that a dispositional account of cares is promising. We have then spent two chapters on how best to understand both dispositions in general and psychological, including emotional, dispositions in particular; emotional dispositions, construed realistically, are things we have good reason to accept. It is now time to complete our account of caring.

Recall the Simple View of cares introduced and partially defended in Chapter 1. According to this view, to care is to be disposed to feel a range of emotions towards a given focus in certain situations. We saw that a plausible but unacceptable way to construe the notion of a disposition to feel is in terms of a pattern of emotions that are partly actual, partly counterfactual. The main problem with this view is that patterns appear neither necessary nor sufficient for the presence of a care. An alternative option, however, is to construe such a disposition realistically, namely as an entity that is distinct from the emotional episodes that constitute its manifestation and therefore as a causal determinant of them. This is what I shall call the realist-dispositional account of cares (RD):

(RD): To care for x is to be disposed to feel for x, where to be disposed to feel for x is to have an intrinsic dispositional property or causal power that, depending on the degree of disposition—‘friendliness’ present in the situation, may produce a range of emotions in the appropriate triggering conditions.

There are several elements of (RD) that need unpacking. First, the relevant sort of disposition is to be construed as a property or causal power distinct from the range of emotions that it produces ‘depending on the degree of disposition-
‘friendliness’ present in the situation’. The addition of the latter clause, admittedly loose but nonetheless (I hope) intuitively clear, is to preserve the idea that, although having a disposition D to A in C makes it more likely for one to A in C, having D does not make it necessary that one will A in C. After all, D could be constantly masked. This is why it is appropriate to say that the disposition may, rather than will, produce certain events. A care, as a result, is a disposition that has possible manifestations but whose existence does not depend on the occurrence of such manifestations. Moreover, a care is an intrinsic disposition, as it can be masked. Although the addition of the clause mentioning other ‘friendly’ factors may suggest that the disposition is extrinsic, it would be a mistake to think so, as the disposition surely exists when such factors are not in play (and, therefore, the factors cannot constitute part of the disposition’s bearer).

Now, it is important to note that the account is ‘minimal’ in the sense that it provides the conditions that must be met for any creature to be correctly said to care about something. Having a certain sort of emotional disposition towards a focus F is both necessary and sufficient to count as caring about F. An advantage of the view is that human infants and certain non-human animals are allowed to care about their parents or their young. The view, however, does not claim that the relevant disposition can manifest itself in emotion only. It is natural to think that caring can also manifest itself in judgement (‘I love her so much’), desire (the desire to buy flowers for one’s beloved) and action (the buying of the flowers). And it is plausible that such events can occur in the absence of emotion. Although emotion can produce desires, judgements and actions—and this is an advantage of the view that it can explain the latter by appealing to the occurrence of the former—it seems possible for the latter events to occur without being produced by an emotion and still count as a manifestation of the relevant care. Moreover, it seems possible that I form the habit of behaving in a certain way towards a loved one (for instance, by buying her flowers every month) and that habit to be expressive of my love for her. What are we to make of such cases? Is (RD) too restrictive in making a certain

---

58 See Chapter 1 for the notion of focus.
sort of disposition to experience emotions both necessary and sufficient for caring?

What I would like to suggest is that (RD) is not too restrictive as it not only allows for other kinds of mental or physical events to be produced by the disposition, but also has the resources to explain why such events can constitute its manifestations. First, although it is not part of the nature of the relevant disposition to produce such events, nothing in the account precludes the disposition from manifesting itself in events other than affective states. The account however does not explain why this is possible and why such events can constitute the manifestation of the disposition. After all, since it is possible to have exactly similar desires, make exactly similar judgements and act in an exactly similar way but fail to care about the relevant object, something must be said about how such events could adequately be connected to the relevant cares. In other words, the realist-dispositional theorist must give us a plausible story of the mechanism by which this could happen.

I suggest that the relevant events must be properly connected to the subject’s patterns of emotions—the ones produced by the relevant disposition—in order for them to count as the expression of a care. Put differently, judgements, desires and actions can count as an expression of caring by being parasitic upon, or grounded in, a certain history of emotional responses. It is only if such responses have occurred, and have played a role in the aetiology of the relevant events that the latter can be said to constitute manifestations of caring.

Here’s an account of how this may happen. When one starts to care about someone, the emotional responses that one undergoes as a result can in turn produce certain desires and judgements (and the actions that can ensue), mental events that only constitute at this stage an expression of the care only derivatively. However, on the basis of a repetition of such events, one may acquire the disposition to form desires and judgements, a disposition that, because grounded in past emotions (and therefore a care), can now be counted as part of the relevant care. So, although it is not necessary for a care to
manifest itself in desire and judgement, it may be the case that a care is (contingently) such that it includes a disposition to form certain desires and judgements.

Questions remain, however, that need to be answered for my account of cares to be complete. First, what distinguishes psychological, including emotional, (intrinsic) dispositions from non-psychological dispositions? So far, all the explanatory work done by psychological dispositions has not been done because they are psychological. Producing mental events is indeed not sufficient for counting as a psychological disposition; things that are not psychological (e.g., a drug) can produce mental events as well. A way certain dispositions to produce mental events may count as psychological will be discussed in Section 3. Second, if cares are emotional dispositions, what distinguishes them from the other kinds? All I have said is that the dispositions necessary and sufficient for caring are emotional dispositions of some sort. As a result, since many dispositions are dispositions to feel a range of emotions, more needs to be said as to what makes some of them cares. This issue will be dealt with in Section 4.

2. The ethical significance of the mental

Before considering the question of what makes psychological dispositions psychological, it may be a good idea to ask why precisely a realist about sentiments should want sentiments to be genuinely psychological or mental. If it turned out that sentiments as the realist construes them fail to count as psychological, we need to say why this fact may at least constitute a problem for (if not a reductio of) her view. I take it that the answer has to do with the ethical significance, broadly construed, of mental states in general, and psychological dispositions in particular.

Mental states are things for which we can sometimes be praised or blamed or held responsible. For instance, the upholding of racist beliefs in the face of overwhelming counterevidence is something for which we regard others as
blameworthy, even when we are aware that giving up the relevant beliefs may not be, at least not entirely, within their control. By contrast, we don’t praise or blame people for being born on a certain day of the week, or for the colour of their hair, or even for having mere neurological conditions (i.e., physical conditions that are not re-describable in psychological terms), even though these are also properties that they possess. Mental states, therefore, are ethically significant in the sense that they are capable of having a bearing on the worth of the person possessing them.\textsuperscript{59}

Mental states are not only capable of having a bearing on the value of the person possessing them; they often have a bearing on the value of certain events in the world, namely our actions. Compare the killing of innocent people by a mass murderer with a thirst for blood and the killing of innocent people by a hurricane. Although the consequences (e.g., the number of deaths) may be the same in both cases, there is an obvious moral difference between the two sorts of events; the event caused by the murderer is immoral whereas the event caused by the hurricane is neither moral nor immoral—it is just very unfortunate. This moral difference clearly depends on the mental difference between the murderer—engaging in purposeful action, having certain beliefs and desires, and so on—and the hurricane—an impersonal entity having no mental life of its own. The fact that certain events are caused by our mental states thus can make all the difference between a right or wrong act and a mere unfortunate event.

Another way mental states can be ethically significant is when they are subject to certain norms of assessment. We assess beliefs as justified or unjustified, desires as rational or irrational, emotions as appropriate or inappropriate, and we can sometimes assess any of these things as morally good or morally bad. We

\textsuperscript{59}We may even go further by saying that mental states, including mental dispositions, are constitutive of the person who possesses them. Compare the disposition of love with an electronic chip that produces loving emotions (Chapter 1). Although both entities are plausibly intrinsic to the person, and therefore can be construed as involving intrinsic dispositions, it is tempting to say that only the former produces emotions that ‘speak for’ the person or that it is a genuine part of the person (as opposed to the organism that constitutes her).
may think that we should strive to be more rational and less irrational, suggesting that we may have some control over the quality of our mental states. Non-psychological entities, by contrast, are not things we describe in these ways. We don’t think of a piece of physical stuff as right or wrong about some issue, or justified or unjustified; we just don’t think of it in evaluative or normative terms at all.

If sentiments construed realistically cannot be psychological, therefore, they cannot be things that we can assess in these ways. They cannot also be things that are capable of having a genuine bearing on our worth or the value of our actions (at least directly). They may even not be able to be part of us as persons (if one distinguishes the person from the organism or animal partly constituting him or her); in that case, they would not be ‘personal’ states. Whether or not the realist about sentiments takes the mental to be ethically significant in all these ways, she cannot deny that ordinary thinking often does and that at least some sentiments—love, caring, long-term jealousy and anger—are ethically significant if anything is. We do think that loving others makes a contribution to our worth as persons, and we do think that a long-lasting anger towards people from our past can be unreasonable. The realist is thus hard pressed either to give an account that is consistent with these ordinary thoughts or to show that these thoughts are mistaken. I think that an account of many sentiments that is consistent with these thoughts is possible, and it will be the aim of the next two sections to propose one for one species of sentiment, namely caring.

3. Psychological dispositions as attitudes

It would be an understatement to say that the question of what makes certain things, and not others, psychological or mental—the so-called question of the ‘mark’ of the mental—has been a major issue in the past decades. What makes a perception of a red vase, the thought that it will be sunny tomorrow, the belief that François Hollande is the current president of France, a pain in one’s ankle, a desire for ice cream, and an episode of sadness because of a break-up all instances of something psychological?
Although there is an on-going debate about what the best answer to this question is, it is instructive to consider the most popular options, as it can give us some idea as to what is likely to be psychological and what is not. I take it that there are two main answers. Let’s consider, first, the view that what makes an entity a psychological entity is that it is conscious. On this rather popular view (e.g., Searle, 1983, Strawson, 1994), what makes a mental state mental is the fact that there is something it is like to be in that state; it is, so to speak, to be part of one’s stream of consciousness. Mental states, on this view, are necessarily phenomenal states. A quite plausible idea, for proponents of this view, has it that being phenomenal is also sufficient for an entity to count as psychological. If an entity has phenomenal states, it is natural to think that it also has a psychology. Although the claim of sufficiency is difficult to deny (although it is not entirely uncontroversial), it is relatively easy to doubt the claim that being phenomenal is necessary for a state to be psychological. Are all mental states things that are felt by the subject having them? Of course, it is hard to deny that many mental states are things we can experience in the relevant way, but is it really the case that all mental states are experiential?

Given the considerations provided throughout this thesis, the answer the realist about psychological dispositions must give is no. Not all kinds of mental states seem to have a phenomenal character. Beliefs and desires, at least in dispositional form, are not things we can feel; all we can feel are the mental events that they produce. If the view under consideration was right, there would be strictly speaking no such thing as a psychological disposition, at least understood as an entity that can exist in unconscious form; the entities doing the causal work that I’ve been talking about (including what I called the emotional dispositions) would thus all constitute brute subpersonal entities that have no role to play in psychological explanations. In this case, we would be provided a reason to think that, although we could be realists about the relevant dispositions, we should be error theorists about the common claim that these dispositions are psychological or mental.
The realist about psychological dispositions (at least construed as entities that can exist in unconscious form) must therefore reject the view that being conscious is necessary for being mental. I take it that there are two main ways to do this, the first ambitious, the second modest. The first way to reject the view is to show that it is inadequate. The realist could for instance point to the fact that common-sense is committed to the existence of psychological states that are unconscious. It is certainly natural to think that beliefs and desires can be unconscious and still count as mental or psychological. The opponent, however, could reply that she is putting forward the claim that common-sense is wrong to think in this way and that the realist begs the question against the error theorist if she points to commitments that the error theorist takes to be false. The realist can then attempt to show that the error theorist’s claim is false, which would force her to engage in the ambitious project of finding the so-called mark of the mental. The second, more modest, strategy is for the realist to help herself with an alternative conception of the mental that is both consistent with the existence of unconscious psychological entities and at least as plausible as the first view. This is the strategy I will now pursue.

Psychological dispositions, we have said, are things that can take a non-conscious form. So the claim that they are genuinely psychological is inconsistent with the theory that claims that being conscious is necessary for a state to be mental. Let’s consider the second option. According to the recent orthodoxy, what all mental states have in common is that they are intentional—they are all about some thing or other. A perception of a blue bicycle is about a particular object, namely a blue bicycle; the thought that all swans are white is about a collection of things, namely all swans; and as we saw in Chapter 1, an episode of joy at the view of a loved one is directed towards a certain object, namely a person that the subject having the emotion loves. At this point, it is

---

60 She can however allow that being conscious is sufficient for being mental or that if a given entity is conscious, it’s necessary or at least very likely that it also has a psychology.

61 This is compatible with the possibility that some mental events involve dispositions that may be felt. These dispositions are however not the psychological dispositions I have been speaking of.

62 For the rest of this section, I am indebted to Tim Crane’s very clear discussion in his 2001, especially Ch. 1 and 4.
natural to conclude that, since intentional states take objects, the latter objects must in fact exist. But this would be mistaken. If there is a fact about intentional states that is established, it’s the fact that at least some of them can be about an object O even though O does not exist, never existed, and will never exist. If one believes in unicorns, there is something that one believes in, namely unicorns, even though there are no unicorns; in this case, one is simply wrong in thinking that unicorns exist. Many intentional states, as a result, have conditions of accuracy or correctness.

Furthermore, intentional states present their object as being in a certain way. “For there are many ways to imagine my mother in her kitchen: she might be baking bread, she might be listening to the radio, she might be frying onions, and so on. A particular episode of imagining my mother will present my mother in one way and not in others. These ways need not be determinate in every respect. But every episode of visual imagining will certainly exclude some ways of presenting the object of the episode.” (Crane, 2008, 4) Saying that a mental state is something that has an object does not, as a result, capture everything there is to capture about the fact that it is intentional. We must therefore introduce the notion of intentional or representational content, roughly the way the mental state presents its object. Intentional states, as a result, are mental states that have representational content.

Finally, each type of intentional state constitutes a distinct ‘mode’ or attitude directed towards a content. I can imagine that \( p \) and I can believe that \( p \). Both cases involve the same content, here the proposition that \( p \), but do not involve the same attitude towards the content: one attitude is imagining, the other is believing. Some philosophers think that all attitudes are propositional in the sense that their content involves a proposition about the world, as in believing that \( p \), imagining that \( p \), and desiring that \( p \). However, it is becoming quite common (e.g., Montague, 2007) to think that it is implausible that all contentful states or attitudes must involve a proposition. Perception, for instance, hardly

---

63 I accept the possibility that some kinds of mental states are ‘factive’ in the sense that the having of one towards an entity \( E \) entails that \( E \) exists. If knowing is a mental state, it is factive. Some people take it that perception is factive as well, a view I do not wish to commit myself to.
seems amenable to being analysed as a simple propositional attitude. The same is true of love: we don’t love propositions; we love people. Furthermore, some have argued that certain attitudes have a content that is partly or wholly non-conceptual (e.g., Evans, 1982, Crane, 1988). Although I have some sympathy with such a view, I will here try to remain silent on whether it is true, as it is not relevant to my larger argument.

Brentano’s original thesis is that all and only mental states are intentional: there is no mental entity that is non-intentional (i.e., intentionality is necessary for mentality) and there is no non-mental entity that is intentional (i.e., intentionality is sufficient for mentality). The conjunction of these two theses is what is meant when we say that intentionality is the mark of the mental. Can intentionality be the mark of the mental? Can it be what makes psychological states psychological? On the face of it, it seems that the answer is negative. The most natural way to challenge Brentano’s thesis is by finding counterexamples to the claim that intentionality is necessary for mentality. It certainly seems obvious that some mental states are not directed at anything in the way that beliefs, desires, perceptions and emotions are. For instance, a gloomy mood, a sensation in one’s body, a general state of anxiety, and a startle response all may seem to be about nothing in particular. One strategy for the defender of Brentano is to argue that even such states exemplify intentionality, albeit in a way that may be different from other mental states. Moods, for instance, have been claimed to be evaluations about the world as a whole (e.g., Crane, 1998, Goldie, 2000). Moreover, sensations, states of anxiety and reflex responses may all constitute primitive forms of intentional states that we may share with animals, an idea not so implausible if we already accept the claim that attitudes can have non-propositional objects (Montague, 2007). Whether or not this solution is successful, it is quite likely that if a thing is intentional, then it is also psychological. What is important in our context is that, if psychological dispositions could be shown to have intentionality, then we have vindicated the claim that they are really psychological. And no strong argument is needed to establish that beliefs are the sort of thing to have intentionality! The fact that they take a dispositional form does not entail anything about their status as an
intentional entity (unless of course we also accept the view, which, as realists, we can reject, that only conscious entities can be intentional).

One difficulty however is that there may be cases of non-mental entities that exemplify something very much like intentionality, such as plants’ disposition to move towards the source of light, suggesting that intentionality is not sufficient for mentality (Crane, 1998). If this is true, this will complicate the matter considerably, as the things whose psychological status we want to establish are themselves dispositional. Fortunately, we need not accept the very controversial thesis according to which non-mental entities can be intentional, or at least can be intentional in the same way that mental entities are. As we have seen, mental states that are intentional have a content, propositional or not, that provides their object, and they are divided into distinct kinds of intentional modes or attitudes that bear different relations to their content such that different kinds of attitudes can be directed at the same content. We can conclude, as a result, that if an entity has a representational content and can be individuated as a distinct kind of mode or attitude, then it is psychological.

Given that beliefs and desires are intentional, we can safely say that they are psychological. And given that beliefs and desires can exist in dispositional form,64 we have good reasons to believe that at least some dispositions are genuinely psychological. One however can accept this much but still deny that emotional dispositions are intentional and therefore claim that they fail to count as psychological.65 If we are to accept the claim that, for an emotional disposition to be genuinely psychological, it must be intentional, then it seems that at least some emotional dispositions may come out as non-psychological. Some propensity towards certain emotions may be an innate part of our personality that has nothing to do with representation (perhaps the so-called ‘temperaments’), or may be the result of a change in brain structure. We can

64 I suspect that belief strictly speaking can exist only in dispositional form. Although one can be conscious that one believes that p, one, it seems, cannot consciously believe that p; at best, one makes the occurrent judgement that p. (See Crane, 2001, 105-108.) The same may not be true of desires, as it seems that I can consciously desire to eat.
65 And, of course, they cannot be shown to be psychological on the basis of their being conscious, which I have assumed they are not.
imagine, for instance, a trait that randomly triggers negative emotions in people who have suffered from a certain brain injury. Although this trait counts as an emotional disposition, as it is a disposition to feel certain emotions, it intuitively does not count as psychological. The trait is akin to a drug, except that it is integrated into the subject’s brain. Consequently, we can accept that certain emotional dispositions are not genuinely psychological.

This claim neither implies that all emotional dispositions are non-psychological nor shows that emotional dispositions, psychological or not, fail to be genuinely explanatory. Of course, if all emotional dispositions are non-psychological, the explanations given in the previous chapter turn out to fail to be psychological explanations. This would be surely a problem, as we have seen that a variety of antirealist about emotional dispositions can readily accept that emotional episodes are caused by non-psychological causal powers in us. The difference between the realist and this kind of antirealist would therefore lie in the fact that, while the realist calls ‘emotional dispositions’ such non-psychological causal powers, the antirealist prefers to use that label for patterns of psychological entities, namely the emotional episodes themselves. The antirealist would as a result have the upper hand, at least if we accept the rather plausible thought that emotional dispositions should be analysed in terms of psychological entities. Calling emotional dispositions non-mental bits of the world hardly does justice to the idea that these things are part of our psychology.

Fortunately, we do not need to accept this conclusion, as some emotional dispositions construed realistically are plausibly psychological. For instance, I have argued elsewhere (Naar, forthcoming) that an account of love as (at least in part) an emotional disposition construed realistically is best in accommodating a number of intuitions, including intuitions on cases of masking. If this is true, we have at least one example of an emotional disposition that is intentional, hence psychological, for love is commonly said to have an object construed under a certain aspect (I love my parents, not the aggregate of atoms that happen to constitute them). We will shortly see (Section
4.2) that caring in general, of which love may be a species, is an attitude that is intentional, and as a result that it is genuinely psychological.

4. What kind of psychological disposition is a care?

I have argued that to care is to have an emotional disposition of some sort. I have however not shown what sort of emotional disposition a care is: although it is on my view necessary, being an emotional disposition per se is not sufficient for an entity to constitute a care. Moreover, I have not shown that the sort of emotional disposition a care is is genuinely psychological. In this section, I will attempt to answer these two questions.

4.1 The centrality of caring

If caring is an emotional disposition, it should be of a special sort. Intuitively, to say that a person cares about something is to say something significant about that person. It is not merely to say that she is likely to feel and behave in characteristic ways, although this may be part of what we mean in certain contexts. It is primarily saying something intuitively deep, rather than superficial, about the person. Contrast “X cares about Y” with “X likes Y”: while the latter suggests a state that may be rather superficial, although it need not be, the former suggests a state that is comparatively deep or central to the person. One might loosely say that, although likes don’t necessarily ‘define’ the person, it seems right to say that cares do. Cares thus seem to differ from other emotional dispositions in the structure they exemplify.

Let’s call centrality the general feature that intuitively distinguishes cares from other emotional dispositions. We now need to say a bit more on what makes certain emotional dispositions central. I will suggest several properties that, taken collectively, contribute to an emotional disposition being intuitively central or deep. Although a list of necessary and sufficient conditions seems hard to come by, I think that if an emotional disposition has all of these

\[66\] Especially when we add the locution ‘really’, as in “X really likes Y”.

129
properties, then it is very likely to be a care. Given the difficulty in individuating cares from other psychological/emotional dispositions on purely structural or formal grounds, I will in the next section suggest that cares should be conceived as a special sort of attitude distinct from the other psychological attitudes.

I suggest that there are at least five properties that mark cares as central emotional dispositions:

First, cares are *intrinsic* emotional dispositions. This by itself does not say much, but it does distinguish cares from extrinsic forms of emotional disposition, if any. It also suggests that a condition for a psychological disposition to be central in the intuitive sense above is for it to be intrinsic to the person that has it.

Second, cares are *multitrack* emotional dispositions, where a multitrack disposition is a disposition that has multiple manifestations. This, again, does not distinguish cares from many other emotional dispositions, but it does distinguish them from ‘single-track’ emotional dispositions, such as one’s disposition to experience fear at the sight of a spider. The idea that cares have multiple manifestations indeed enjoys some degree of plausibility. Recall Shoemaker’s remark in Chapter 1 that caring involves a ‘package deal’. Although we have rejected his reductive account on the grounds that having a care doesn’t necessitate the feelings characteristic of caring, we can grant him that caring is the *kind* of thing that make us feel in a great variety of ways. Caring, further, might have been less important to us if it rarely produced such feelings. Claiming that caring doesn’t necessitate its manifestations in the relevant conditions is compatible with claiming that it is the kind of thing that, if all goes well, will manifest itself.68

---

67 In addition to that, we can safely say that cares also have multiple *triggering conditions.*
68 One may worry that I am here resorting to a ‘ceteris paribus’ account of dispositions, an account that I have rejected in Chapter 2. This would be misunderstanding the aim of the present discussion, however. *At this stage,* indeed, we are not after an account of dispositions at all, but after a way to demarcate certain (intrinsic) dispositions from others. The appeal to an ‘if all goes well’ clause is not to provide a reductive analysis of any sort but to express a pre-theoretical
A third property of cares may be their resistance to change. A care is the kind of thing which we can’t easily get rid of. If love is indeed a species of care, or at least involves a care, it provides a good example of something that we can only give up over an extended period of time.⁶⁹ Although this does not in itself distinguish it from certain emotional dispositions that are not cares (such as phobias), it clearly goes towards explaining why cares are central.

A fourth way to demarcate cares from most other emotional dispositions is by appealing, not to the sheer variety of emotional responses that can constitute its manifestations, but to the nature of these responses. Many emotional episodes experienced as a result of caring themselves seem to be deep in a way that other emotional episodes may not be. As we have seen, if X cares about Y, X is disposed to grieve if Y were to die. Grief, moreover, is experienced as something of extreme significance, and cannot be so easily discarded as an occasional annoyance. If one grieves over the loss of a loved one, this is partly because one’s life is radically changed—changed for the worse—and grief is evidence that one is aware of that fact. This is not, to be sure, like having a mild sense of discomfort, as when one is standing on a balcony, aware of the void. This is the kernel of truth found in the pattern account introduced in Chapter 1: to care is to be disposed, not merely to feel a range of emotions, but to feel, minimally, a range of emotions of specific types.

The last property that must be emphasised is the capacity of cares to explain the existence of other mental states. First, given the variety of mental events that a care can produce (and this includes the contingent judgements and occurrent desires discussed above), cares can explain events that would not have happened in their absence. Of course, mimicking cases, we have seen, are metaphysically possible, but it is clear that, in normal situations, namely in situations where the patterns of mental events are produced by the relevant care,

---

⁶⁹ At any rate in the paradigm cases. I do not accept the claim that it is conceptually impossible for a love to go out of existence without resistance. After all, the death of someone who loved his wife arguably implies the death of the love.
cares explain their occurrence. Second, cares may explain the existence of other psychological dispositions, including other cares. If I care about justice, I may also care about equality between genders, given that justice may require equality between genders. The point is not that, if I care about justice, I will also care about equality between genders—I may after all think that the two are unrelated. The point is that, in the case where I have both cares, one may explain the presence of the other. Cares, as a result, may come in the form of an interconnected network of dispositions. In addition, cares may explain the existence of other kinds of psychological dispositions. For instance, the fact that I care about someone rather than am indifferent to them can explain why I form various flattering beliefs about them.

These are intuitive ways to mark out cares as distinct from other psychological dispositions. But this overlooks a crucial question: are cares genuine psychological entities?

4.2 Caring as attitude

So far, I have not shown that cares constitute genuine psychological entities. To do so, I must show that cares can plausibly be construed as attitudes of some sort. We have seen that it is hard to deny that love is an attitude that is directed at an object viewed under a certain aspect. So if love is a species of caring, then we have an example of a care that is genuinely psychological.

Can we generalise this conclusion to all cares? Can we show that to care is to have a certain attitude towards the world? I think that the answer should be obvious by now. Cares are attitudes we take towards the world. The precise nature of these attitudes, however, is the difficult issue, and will be tentatively tackled in Chapter 6.

70 One may think that this is a clear case of an undue influence on belief formation. However, as will be clear in Chapter 6, this partly depends on the kind of attitude we take cares to be.
In order to establish that cares are attitudes, it is natural to start with language use. When I say that ‘I care about my parents’, there is something that I care about, namely my parents. The same is true with the locution ‘care for’: caring for is always caring for something. Caring has therefore the property of being directed towards the world, of having an object. Next, as suggested earlier about love, caring is always caring about something as viewed under a certain aspect. In the example given above, caring about one’s parents is not the same thing as caring about the matter that makes up one’s parents, even though the two may turn out to be identical. Moreover, I can care about the health of my parents even though I may not care about the specific biological facts that constitute being in good health. As a result, caring about things in the world is always caring about these things as represented in a certain way.

Nothing I have said so far implies that cares are propositional attitudes, however, or that they necessarily involve a content that is wholly conceptual. It is an open question whether these claims are true of cares. It could certainly be the case that cares involve some form of non-conceptual content, or that some cares are propositional while others are not. It is quite clear, however, that many paradigm cases of cares are not, on the face of it, propositional. We don’t often care about propositions or states of affairs; many times, our cares are directed at particular entities, people being the prime example. Yet, at the same time, it seems that some of our cares may be propositional. If dispositional desires can constitute instances of caring (Ch. 3 & Ch. 6), then since desires can take a propositional form (e.g., ‘I desire that X is the case’), cares can take a propositional form as well.

This leads us to the question of what kind of attitude caring could be. This question is harder to answer than the question whether cares are attitudes of some sort. Several possibilities are open, but two suggest themselves. On the one hand, a care may be a special case of belief, namely a belief that is central in the sense given above. On the other hand, a care may be a special case of dispositional desire, namely a ‘central’ desire. It is important to note however that the claim that cares are either beliefs or desires seems to imply the denial of
an advantage that the realist-dispositional account of cares initially had, namely the fact that the account allows creatures that may not be capable of forming propositional attitudes to be genuinely capable of caring. If to care is to have a certain kind of propositional attitude, then the appeal to emotional dispositions in order to avoid intellectualism loses its rationale.

In the rest of this thesis, I will assume that beliefs and desires do not exhaust the range of cognitive and conative attitudes that one can have, as some forms of attitudes, such as an infant’s attachment to his mother, can take a non-propositional form. The question will therefore be whether cares should be conceived as broadly-speaking cognitive, conative, both, or neither, whether they are reducible to other kinds of entities or are sui generis, and whether the form that cares take can constitute a suitable object of normative assessment. Although a complete answer to this question is hard to come by, I will argue in Chapter 6 that at least some cares (and more generally sentiments) are partly cognitive and are subject to epistemic assessment.

4.3 Filling in the details of the realist-dispositional account

Until then, let’s take the attitude of taking as valuable as a placeholder for the sort of attitude(s) that a care is. I intend this formulation to be neutral between caring being cognitive and caring being desiderative or conative (or both, or neither), and between caring being propositional and caring being non-propositional. What is important to accept is that to care about something is to take certain things in the world (as they are or as they should be according to the attitude) to be valuable to oneself.

This allows us to give more flesh to our account of caring. On the realist-dispositional account I propose, to care about something is (a) to be disposed to feel a range of emotions towards it. This disposition to feel, moreover, (b) is central to the subject’s psychology and (c) constitutes an attitude whereby the object of the care is presented as valuable to the subject. Less informally:
Caring for X is the central attitude of taking X as valuable that disposes one to feel a range of emotions towards X (or to things suitably related to X).

Of course, the crucial question now is, what does that attitude of taking as valuable consist in? This question won’t be tackled until Chapter 6. However, it should be noted that there appears to be no easy way to reduce cares to well-known entities such as beliefs and desires, given that creatures capable of caring may not be capable of acquiring propositional attitudes (Chapter 1). The question we will ask, as a result, is whether caring should be conceived as a cognitive or non-cognitive state, broadly construed; should cares be, in other words, construed as states purporting to represent the world as it really is (having thereby a mind-to-world direction of fit) or should they be conceived as representing the world as it should be (according to the attitude they constitute)? We will see in Chapter 6 that at least some forms of caring involve a cognitive component.

5. Caring and character

In this section, I would like to show that caring, as understood by the realist-dispositional account, can be naturally connected to the notion of character. The idea that to have a character is, at least in part, to care about certain things is not original. It is natural to think that virtues, a species of traits of character, require the possession of certain cares. The honest person cares about the transmission of truths and the absence of deceitfulness. The generous person, at a minimum, cares about the wellbeing of other fellow humans (and perhaps non-human animals). On my account, this implies that one is minimally disposed to feel a range of emotions in the situations that are relevant to the trait—such as when one sees someone who needs help, witnesses an act of generosity in others, realises that one failed to do the right thing in a given situation, and so on. This is quite intuitive: if one fails to feel guilty when one knows that an act of generosity was called for and that the kind of person who is generous would have performed the act, we have evidence against the presence of the relevant virtue. A vicious person is someone who either cares about the wrong things or
simply fails to care about the right things. A cruel person, for instance, may care about the suffering of others or at any rate fail to care about their wellbeing. These simple points suggest that there is a strong connection between cares and character. In this section, I’ll make an intuitive case for the view that cares as I understand them are central to character. In particular, I’ll suggest that the two are conceptually connected (i) on the dimension of disposition and (ii) on the dimension of attitude.

First, we need a characterisation of what it is to have a character. I intend the following list of features to be as theory-neutral as possible:

a) Traits of character are dispositions of the mind that have characteristic manifestations, including feelings and actions. Although behaviour is the main source of evidence of the presence of traits, it cannot by itself conclusively show that one doesn’t possess a given trait, at least at a given time and for certain traits. For instance, the fact that a social worker once fails to show compassion to someone in need does not conclusively tell against her having genuine compassion.

b) Appealing to character traits serve certain explanatory roles. There is a use of the notion of character which is not evaluative or normative. It is possible to extract from the sentence ‘This man is honest’ a purely descriptive claim such as ‘This man is disposed to so-and-so in such-and-such circumstances’.

c) Character traits are those traits that, although they need not be morally significant at every moment at which they exist, could be morally significant. This is to allow character traits both to become virtues or vices and to be neither virtues nor vices at a given time. Moreover, character traits are the kinds of thing that can have a bearing on the worth of the person possessing them. They

I make the common assumption that virtues and vices are traits of character. See Hurka, 2006, for an alternative position, as well as virtue reliabilists such as Ernest Sosa and John Greco.
differ from mere capacities and skills, things for which we do not typically admire others in the relevant personal way (Baehr, 2011).

d) Character traits differ from mere personality traits, although they might form a subset of them—that of the potentially morally significant. Both personality and character trait ascriptions are often meant to individuate people. If I say that my neighbour is kind, generous, but lazy, my claim implies that he differs from the people who are unkind, those who are ungenerous and those who are industrious. And the longer I make the list, the less people will have the same kind of character as my neighbour’s.

e) Having a character and having particular character traits may be different concepts. One’s character seems related to one’s identity, and seems to be able to manifest itself in countless situations—it is like a global condition of the mind. One can be a sceptic about particular kinds of character traits (e.g., honesty, generosity, etc.) without being a sceptic about character in general. However, not vice versa: if one has traits of character, one thereby has a character. There is a question of how to type-individuate traits of character. There may be many such traits that common-sense doesn’t recognize (see Chapter 5).

Now, I would like to suggest that traits of character both are central dispositions and require certain cares. In favour of the first claim, we can point to the facts that character traits are plausibly intrinsic and multitrack, that they are typically resistant to change, and that they can explain both the existence of certain events (e.g., feelings, desires, intentions, actions) and the existence of other psychological dispositions (a generous person tends to have certain dispositional desires and beliefs about the world). Character traits, as a result, seem to have a structure that is similar to that of cares. In favour of the second claim, we can point to the fact that a character trait is not merely a disposition to behave in characteristic ways; it is a disposition to behave in characteristic ways in a
world towards which we have certain attitudes. What distinguishes a genuinely virtuous person and a machine that merely mimics virtue is that the former’s actions are related in some way to what the person, broadly speaking, thinks about the world. I contend that at least part of what she thinks about the world is constituted by what she cares about.

I’ll remain silent on whether caring can be sufficient for character. One possibility is that it can be, but that for traits of character to become virtues, certain further conditions must be met (e.g., Baehr, 2011). Another possibility is that character traits themselves require also a cognitive or behavioural disposition such that emotions can be bypassed. But my account of cares, as we have seen, may accommodate such cases by making the relevant events (and the dispositions to these) parasitic on a prior history of emotional episodes. Although this may be the right way to go about character traits, it may seem implausible when it comes to those character traits that are also virtues, as it would have the counterintuitive consequence that having a disposition to certain emotions, as opposed to action, is sufficient to instantiate virtue. This, indeed, would have the consequence that, as long as one had compassionate feelings, one could perform the nastiest things and still count as a compassionate person. I conclude that, although cares are necessary for character, they may be insufficient, at least in the case of virtuous character traits. I will resume the matter in Chapter 5 when I will propose an account of the relationship between character traits and virtues.

Despite what has just been said, one may still be sceptical of the idea of having a character. A recent wave of empirically-informed moral psychology has it that character traits either do not exist or are so rare that they shouldn’t have such a prominent role in contemporary ethical theory. In the next chapter, I will give reasons to be wary of this trend. Although I won’t presuppose the view outlined here, it will be clear that what I say there can complement it.
Chapter Five

The Reality of Character: A Reply to Situationism

1. Introduction

In this chapter, I will be concerned with a recent wave of scepticism about character traits in general, and virtues and vices in particular, known under the name of ‘situationism’. If we are to trust a widely held interpretation of the empirical literature on the matter, people either rarely possess temporally stable and robust character traits (Doris, 2002) or do not possess any at all (Harman, 1999). Moreover, data suggest that, even if character traits existed, we would be pretty bad at acquiring knowledge, perhaps even justified belief, about them (Alfano, 2011). And given that our practices of attribution of traits of character are so unreliable, there may be no good way for us to have evidence for their existence, which may justify outright antirealism about character traits (Harman, 1999, 2009).

The situationist usually places herself in the position of a challenger: to anyone who believes in the relevant traits, and especially to anyone whose theoretical commitments require the existence or plenitude of such traits, the situationist asks for evidence that realism about them is in fact warranted, and that her alternative hypothesis (that traits are rare or non-existent) is therefore false. But given that, the situationist thinks, the evidence available is straightforwardly evidence against the existence or plenitude of traits, there is pressure on the realist to give up her commitment, and everything that depends on it. This is supposed to be the fate of the virtue ethicist who works within the Aristotelian tradition; as she believes in the kind of traits that the situationist data show do not exist (or are, at best, very rare), namely character traits (and especially virtues), she ends up with the unpalatable thought that people should acquire
traits that are unobtainable. If ought indeed implies can, then virtue ethics, at
least in its Aristotelian guise, cannot be accepted as a normative theory, perhaps
giving way to other kinds of virtue theories that lack the relevant commitment
(e.g., Merritt, 2000).

In fact, it is not just realism about character traits that must be shown to be true;
the situationist who believes that character traits exist but are rare presumably
counts as a realist. The kind of realism at issue here is the one that claims that
character traits not only exist, but also are widespread in the population. If
virtue ethics is to be an egalitarian theory, and even though people may not all
have virtues, it should be the case that most people possess the very material
that may enable them to acquire virtues, namely character traits. Or if most
don’t have them, they could acquire them without too much difficulty. So
realism about character traits here is meant to capture two claims: that (i)
character traits exist, and that (ii) if character traits are not widespread across
the population at the moment, which is prima facie implausible, minor changes
in the world should lead to this result. What I will call ‘strong situationism’
denies (i) and what I will call ‘moderate situationism’ denies (ii). (What I will
call ‘weak situationism’, by contrast, only claims that appeals to particular
character traits in everyday explanations are, at least many times, unfounded, a
view that, I will suggest, the virtue ethicist can accept.)

In this chapter, I will argue that both of these forms of situationism, by helping
themselves without argument to a controversial methodological assumption,
makes the situationist’s case against the virtue ethicist easier than it should be.
The assumption in question is that evidence against particular virtues
(generosity, compassion, honesty, cruelty, and so on) is ipso facto evidence
against character traits in general. If the empirical data suggest that a virtue is
not at the origin of a given piece of behaviour, they also suggest, on this
assumption, that no character trait—and what can be called ‘character’ in
general—can be at its origin either. I’ll argue against this assumption by
suggesting that the account of character the situationist works with fails to
capture the dispositional nature of some central entities accepted by ordinary
thinking about character and differing from the dispositional nature of the virtues. I’ll then provide the outline of an account of character traits, and of their relationship with particular virtues, that the virtue ethicist can help herself with in responding to the situationist challenge.

One note on methodology is important to give. In this chapter, I will, for the sake of argument, grant to the situationist as much as possible. For instance, I’ll grant the claim that the data show that many of our attributions of traits are based on biases and that they are therefore in fact mistaken a lot of the time. I’ll also grant that the situationist data show that many virtues are at best rare, which may make it look like I am in fact on the situationist’s side. My claim however will be that even granting these claims, it is possible for the character realist to provide a persuasive case for her realism, and for the virtue ethicist to argue that we should still be optimistic about our prospects of acquiring virtues.

2. From thick concepts to empirical predictions

Before introducing the various forms of situationism, and the kinds of data on which they rely, it is important to have a clear grasp of the situationist’s methodology. Although it is rarely made explicit, here’s a way of setting out the situationist’s central argument. The situationist’s initial reasoning can be reconstructed as following three basic steps. First, terms and concepts of the disputed class of psychological entities are identified. The disputed class mainly includes virtues, such as compassion and honesty, and vices, such as cruelty and stinginess. These terms and concepts are ‘thick’ in that they have both a descriptive content and an evaluative content. The second step consists in making the relevant concepts ‘empirically manageable’ by extracting (at least some of) their descriptive content. This may be done by formulating plausible platitudes such as ‘If someone is courageous, she is likely to stand her ground in the face of opposition’. On the basis of the results of the second step, a definition, that captures the terms and concepts of the disputed class (those identified in step 1), is formulated—this is step 3. This definition must, at least in part, roughly correspond to the one given by the virtue ethicist, and need not
involve conditions that are necessary and jointly sufficient. Several definitions may therefore be put forward, but one general feature that is often appealed to is that character traits are persisting dispositions of the mind that are reliably manifested across situations of both the same type (stability thesis) and different types (consistency thesis) (Doris, 2002; see also Harman, 1999). Honesty, for instance, is the kind of thing to manifest itself not only at work, but also at home, on the streets, with friends, with strangers, with family, when in a hurry, when at leisure, and so on and so forth. If one only tells the truth when one is at work, or, more dramatically, if one only tells the truth to one’s boss, this is evidence that one fails to have the character trait of honesty; one may at best have the local trait of honesty-at-work or, worse, honesty-when-one’s-boss-is-present (Doris, 2002), hardly the kind of trait that the virtue ethicist wishes to instil into people. The situationist typically claims that the thought that character traits are ‘global’ rather than local is both a commitment of Aristotelian virtue ethics and a commitment of common-sense.72 The issue then is to find out whether anything in the world corresponds to the rough definition of character traits formulated in step 3. The situationist claims that nothing (what I’ll call ‘strong situationism’) or very few things (what I’ll call ‘moderate situationism’) do correspond.

Reasonable disagreement can be found at every step. First, one can disagree about the items of the list, thinking that some item(s) must be added or removed. For instance, one can claim that concepts other than concepts of traditional virtues and vices must be included (e.g., wit). In the case where items are added, this may result in a modification of the definition formulated in step 3. Second, one can disagree about the method of extraction done at the second step. If, for instance, one is a non-reductive naturalist about value properties (e.g., Sturgeon, 1988), thinking that they are perfectly natural properties of the world on a par with ordinary natural properties and objects, one is likely to think

72 I take the claim that character traits are global to be identified with the conjunction of the theses of stability and consistency. One further thesis formulated by Doris, which I will put aside, is that character traits are ‘evaluatively integrated’ (the having of character traits of certain types entail the having of character traits of some other types). I put it aside given that it is controversial whether it is an essential commitment of either virtue ethics or common-sense (Doris, 2002, 23).
that putting aside the evaluative content of the relevant concepts turns out to be putting aside some essential descriptive content. Investigating virtues empirically without working with a conception of them as excellent or at any rate good may therefore be changing the subject. Fortunately, we need not enter this complex issue. If the definition of virtues that the situationist works with does not make any reference to their value properties, it still makes reference to some set of properties that may be essential to them, and if it turns out that even this set of properties is rarely instantiated in the world, then we have reason to think that virtues are rarely instantiated either. So even if the situationist is not, strictly speaking, investigating the virtues—but something very much like the virtues—the empirical data that she appeals to may nonetheless cast doubt on their existence. Furthermore, as will be clear later in this chapter, there is a good reason not to appeal to value properties in a definition of character traits, namely that character traits are not all virtues, and include vices and probably traits that are neither virtues and vices. A general definition abstracting away from the details of specific kinds of traits is therefore needed for the situationist to show, not only that virtues do not exist or are rare, but also that character traits is general do not exist or are rare.

Lastly, one can disagree about the definition given in step 3. For instance, one can claim that it fails to capture all the cases either initially included in the list in step 1 or that should have been included. I will assume that the situationist’s definition introduced above of character traits as dispositions to reliably behave in certain ways is right about the cases that she typically appeals to. It is right, for instance, about compassion, honesty and generosity. I will therefore agree (at least for the sake of argument) with the situationist that the data strongly suggest that at best very few people are endowed with robust traits such as compassion, honesty and generosity. However, I will argue in Section 5 that the definition fails to capture concepts that should be included in the initial list, and should be revised accordingly. These are the generic concepts of character trait and character. Although their inclusion in the list may seem harmless, as it doesn’t imply that the situationist is wrong to think that virtues are at best rare, I will suggest that it opens the door to a strong line of response against the
situationist’s argument for the rejection of Aristotelian virtue ethics. It will turn out that even if the situationist is right about the (current) ontological status of the virtues, the virtue ethicist may still be able to defend a form of realism that is compatible with the data, weakening situationism to the point of being rather uninteresting.

3. Varieties of situationism and implications for virtue ethics

In this section, I will present in more detail what I take to be situationism and what I take to be the situationist’s object of criticism. Although virtue ethics comes in a variety of forms, one of its central commitments is that most of us possess a number of causally efficacious character traits, and, depending on the specific theory one accepts, some of them count as virtues (e.g., honesty) and some of them count as vices (e.g., cowardice). Character traits are here typically defined as dispositions to both undergo a range of mental states (emotions, desires, and so on) and exhibit a range of behaviours across various types of situations. Another commitment is that most of us have the ability to acquire certain virtues, or virtuous character traits, over time (e.g., Annas, 2011). Given that the virtue ethicist’s conception of character traits has a behavioural component, it seems suitable to empirical testing—if we think that someone has a given character trait, or disposition to perform X-Y-Z is circumstance of type C, then it seems that a good way to test our belief is to put her in C and see whether she performs one of the behaviours that we think are relevant to the trait in this kind of situation. More importantly, if we think that most people possess a given character trait, we can test whether this is true by observing their behaviours in the relevant circumstances and see whether they are plausibly relevant to the trait.

This is exactly what a number of experiments in social psychology have been interpreted as doing. Hundreds of experiments suggest that many of our actions are better explained by situational factors having an impact on our occurrent mental states, including our moods, intentions and desires, than on an allegedly deeper source such as character traits. For instance, as Isen and Levin’s ‘dime
experiment’ shows, people are significantly more likely to help someone in need on the streets after finding a coin in a phone booth (14 to 1) than when no coin is found (1 to 24), the difference being explained by the inducement of a good mood in the first group of subjects (Isen & Levin, 1972). Another experiment that is often cited is Darley and Batson’s ‘Good Samaritan’ experiment, where people in a hurry are found significantly less likely to help someone in need (10% helped) than people who are not in a hurry (63% helped) (Darley & Batson, 1973). Lastly, the famous Milgram experiments are often cited as evidence that most people are ready to do whatever an authority figure asks, even such atrocious things as administering high electric shocks to another person, as long as the authority figure is sufficiently close to the subject (Milgram, 1974). In all these cases, it’s situational factors, coupled with occurrent psychological states and processes—mood change, level of hurriedness, etc.—that seem to provide the best explanation of people’s behaviour, rather than character traits.

Additionally, since the results of such experiments fail to match our expectations, it turns out that we tend to overestimate the internal aspect of people (e.g., by attributing character traits to them) and underestimate external factors; we make the so-called ‘fundamental attribution error’ (e.g., Jones & Harris, 1967). Moreover, the presence of several biases and thought tendencies suggests that we would believe in character traits even if they didn’t exist (Alfano, 2011).

Now, various sceptical conclusions could be drawn from these data. First, given that the data are best explained by appeal to entities that are not character traits, we should be sceptical that character traits exist:

Strong Situationism: We have no empirical evidence for the kind of character traits virtue ethicists believe in; therefore character traits do not exist. (Harman, 1999)
A more moderate view is that, although the data do not conclusively show that character traits do not exist, they at least show that, if they existed, character traits would be extremely rare:

Moderate Situationism: Character traits, if they exist, are extremely rare, and therefore people are wrong most of the time they attribute them to others (given that they do it quite often\textsuperscript{73}). (Merritt, 2000, Doris, 2002, Harman, 2009)

An even weaker position is silent on the ontological status of character traits, but makes a claim about our ability to know about them if they existed, the motivation being the existence of phenomena such as the fundamental attribution error:

Weak Situationism: Character traits, if they exist, cannot be reliably tracked by means of one or more of our psychological faculties, and therefore we are in potentially massive and systematic error about their distribution in the world. (Alfano, 2011)

The implications for virtue ethics vary depending on the form of situationism that one accepts. If strong situationism is true, (Aristotelian) virtue ethics should be given up, as it commits itself to the existence of entities that do not actually exist (and therefore that cannot be acquired). If moderate situationism is true, virtue ethics as a normative theory is unrealistic and should be given up on that basis for a more empirically adequate theory. In contrast, if weak situationism is true, virtue ethics may still be defended, even though the pursuit of virtue may turn out to be an uncertain business.\textsuperscript{74}

\textsuperscript{73} The addition of this qualification is important, for nothing in moderate situationism implies that we cannot sometimes get it right. If we could learn how to detect those rare instances of character traits, and if we could refrain from attributing traits in all situations except in those rare situations in which character traits are at play, then we would stop being wrong most of the time we attribute character traits to others.

\textsuperscript{74} Of course, this depends on the version of weak situationism we accept. Presumably, the virtue ethicist cannot accept a global scepticism about our ability to know (or to have justified belief about) the distribution of particular character traits. There must be some way of gathering evidence for the presence of character traits; otherwise the acquisition of virtues would be akin to the acquisition of things that we can’t possibly know we have—not something the virtue ethicist would be happy to accept. The epistemology of character traits, it will turn out, is an
What distinguishes strong and moderate situationism from weak situationism is the fact that they make an ontological claim about character traits: they either outright reject character traits as non-existent or make a claim as to how few character traits in fact are. Weak situationism, by contrast, is only committed to an epistemological claim about our ability to have knowledge about character traits, and therefore is compatible with antirealism, realism and agnosticism about character traits. It can also be combined with a realist view of character traits that claims that character traits are plenty, or that they are few. I will put aside weak situationism for now, postponing it to the last section of this chapter, and will focus on both strong and moderate situationism (for now, simply ‘situationism’).

The issue, therefore, is whether the data genuinely support the ontological conclusions of the situationist (strong and moderate), whether it can only lead to some scepticism about our epistemic capacities (weak situationism), or both. I will assume that some form of weak situationism is supported by the data on attributions of traits, with the existence of biases such as the fundamental attribution error. The question is whether, in addition to weak situationism, there is good reason to be a strong or moderate situationist.

To see whether the data by itself can support situationism, I suggest, we need more detail about the working definition given at the outset. We saw that character traits are dispositions that are reliably manifested. But what is a disposition, according to the situationist? And how does having a disposition

---

One might think that I overlook the possibility that strong and moderate situationism have an epistemological component as well, suggesting that the distinction between these forms of situationism and weak situationism is less sharp than it appears. I surely do not deny that, e.g., strong situationism can be defended on the basis of epistemic considerations—this is presumably what Harman does (1999, 2009). My claim is rather that, while strong and moderate situationism are essentially theses about what character traits there are, weak situationism is essentially a thesis about our ability to know (or have justified belief) about the presence of character traits if they existed.
relate to manifesting it? In the following, I’ll challenge the situationist into providing an adequate answer to this question that succeeds in securing her ontological thesis formed on the basis of the data. I’ll discuss two possibilities. The first will be rejected on the ground that it contradicts thoughts about virtues that the situationist plausibly accepts. The second will be rejected on the ground that it ends up overly restrictive. I’ll argue that, on a more plausible account of the relationship between character traits as dispositions and their manifestations, the situationist fails to show that character traits are non-existent or rare (this doesn’t mean however that character traits exist and are plenty—further argument will be needed). Furthermore, I’ll suggest that the virtue ethicist can adopt my account without significant cost in order to argue for a form of realism about virtues.

4. Character traits reduction

Recall the way the situationist argument is supposed to work. If having a given character trait is to be disposed to perform actions of types X-Y-Z in circumstance C, then we have strong evidence for the absence of the trait if no action of the relevant types is performed in C. And if most people fail to perform any action of the relevant types in C, this is strong evidence that most people do not have the trait. Moreover, if actions of the relevant types are in fact performed (e.g., the people that did help in the phone booth study), the fact that they are best explained by the presence of irrelevant situational factors (e.g., the finding of a dime) casts doubt on the thought that such actions were the expression of the relevant traits. An assumption that seems to be underlying this reasoning is that careful observation of other people’s behaviour at a given time can give us reliable information about their psychological dispositions. We have seen that the situationist data suggest that none of the relevant dispositions are at the origin of most people’s behaviour. The situationist goes even further by making a negative explanatory claim: that character traits do not play a role (or a very limited one) in the explanation of people’s behaviour. In order to secure this claim, however, the situationist must be committed to a view of psychological dispositions that makes her negative explanatory claim plausible.
She must subscribe, for instance, to a view of psychological dispositions as things that can be reliably observed, at least indirectly and with the right experimental machinery. In addition, she must subscribe to a view of character such that, if someone A with a given character trait T finds herself in a situation that is relevant to T, A will manifest T. Or, if the situationist wishes to allow occasional lapses in behaviour, A should regularly manifest T in repeated trials of the relevant situation. In this section and the next, I’ll consider two conceptions of character traits as dispositions that may help the situationist secure her alternative explanation of the data.

A natural way to characterise the relevant psychological dispositions is to reduce them to their manifestations such that, if the latter do not occur, we have conclusive evidence that the relevant dispositions are absent. On a popular version of the view, facts about dispositions are to be reduced to facts about how the relevant entity would behave in a range of situations.\textsuperscript{76} If one were honest, for instance, one would behave in ways that are relevant to honesty in a range of situations where honest behaviour is called for. Even though the absence of honest behaviour in one of the relevant situations may suggest the absence of the trait of honesty, it doesn’t conclusively show that the person is not honest, for she may behave honestly in most, but not all, of the relevant situations and still count as genuinely honest. Determining whether someone (or an entire population) has a given trait, therefore, we must check whether regularity in behaviour is present over the range of situations that are relevant to the trait. If such regularity is present, then the person (or population) arguably possesses the trait.

If this view of psychological dispositions is right, then it seems that the situationist is entitled to make her ontological claim. Given that (let’s assume) we do not observe the regularity relevant to the trait of compassion, at least at the level of the population, over a range of situations that are relevant to the trait, we can be confident that at least most people fail to have compassion.

\textsuperscript{76} This is roughly the ‘simple conditional analysis’ as found (and widely criticised) in the literature on dispositions. See, for instance, Lewis, 1997, Mumford, 1998, Molnar, 2003, and Chapter 2 of this thesis.
Unfortunately for the situationist, the view under discussion is an implausible view about psychological dispositions. As noted by many,\(^77\) the relation between dispositions and their manifestations cannot be that of identity or constitution. At least in many cases, dispositions are thought to explain the occurrence of the events that make up their manifestations. For instance, we commonly say that the reason why a particular object broke is that it is fragile. Moreover, it seems that in the particular case of character, character is thought to causally produce behaviour. Causation is however typically thought to be a relation between distinct existences; a thing cannot cause itself (or at any rate it requires some argument to show that it can). As a result, if character traits were to be endowed with causal powers, they should enjoy some fair degree of independence from the behaviours that they are expected to produce.

Worse still for the view, it seems that even the situationist would not want to subscribe to it, no matter how much it can help her secure her ontological claim on the basis of the available data. Indeed, the situationist could well accept the possibility of a state of affairs where people’s behaviour is both stable and consistent and yet not the production of any trait of character. If situational factors exert the overwhelming influence that the situationist claims they do, then the environment can be engineered in such a way as to make people behave as if they had the relevant traits (Merritt, 2000, Alfano, 2013), suggesting that mere regularity in behaviour, for the situationist, is insufficient for the possession of a character trait.\(^78\) The situationist, therefore, cannot accept the view of dispositions as identical to certain patterns of events.

\(^77\) E.g., Wollheim, 1999, in the particular case of psychological dispositions. See also Chapter 2.

\(^78\) In the literature on dispositions, the relevant case would count as a case of ‘mimicking’. See Johnston, 1992.
5. Character traits as ‘tendencies’ (realistically construed)

So, even for the situationist, character traits must enjoy some degree of independence from their manifestations. And given that she plausibly takes them to causally produce their manifestations, she must take character traits and their manifestations to be distinct existences. But if the two are distinct, then it seems that character traits can in principle exist without ever manifesting themselves. More importantly, it seems to imply that the role of situational factors appealed to by the situationist may not just be to produce behaviour, but to do so by preventing or facilitating the manifestation of the subjects’ character traits. This would put the situationist in a position where she has to defend her negative claim—that the fact situational factors have such a strong influence on people’s behaviour shows that people lack character traits—in the face of an equally plausible explanation if we accept the view that character traits are distinct from their manifestations—that it is because situational factors obscure their dispositions that people fail to manifest them (or, in the case where people do perform trait-relevant behaviour, it is because situational factors enable the dispositions to manifest themselves that they behave in the relevant way). If this is right, then the situationist would need an independent argument to secure her ontological claim.

One such argument is that, even though the dispositions of character and their manifestations are distinct existences, they nonetheless must be reliably connected. As we noticed above, the idea of a compassionate person that never helps other people, or even feels something for them, hardly sounds coherent. And this seems to be true for many virtues. Someone who never gives money or buys anything to others, and who is always grumpy and angry with others, is certainly not a generous, kind and, for that matter, friendly person. In short, virtues—one might say by conceptual necessity—are the kinds of thing that must make some significant contribution to behaviour, at least sometimes. Having a particular virtue, the platitude goes, implies a significant difference in behaviour. Let’s call the view that character traits, including virtues, are both distinct from, but nonetheless reliably connected in this way to their
manifestations the ‘tendential’ view. On such a view, a character trait is a disposition that tends its bearer towards some range of behaviours, where the notion of ‘tending to’ is thought to be a success notion (i.e., if X tends to perform Y, this implies that X has performed Y on a suitable range of occasions already); a trait that lacks this tendency, thus, does not count as a character trait.

How good is the tendential view? As I said, it seems to capture many virtues such as generosity, kindness and compassion. Moreover, it seems to allow a range of frequency of manifestation, depending on the virtue at issue. Mark Alfano (2013) makes a distinction between low-fidelity virtues and high-fidelity virtues.\(^79\) Whereas low-fidelity virtues (e.g., charity, friendliness, generosity, mercy, tact) need not manifest themselves in all the trait-relevant situations, high-fidelity virtues (e.g., chastity, fairness, honesty, trustworthiness) require “near-perfect consistency”. If someone is honest or chaste only 80% of the time, Alfano says, this hardly makes her honest or chaste. By contrast, if someone gives money even 20% of the time, that already counts towards considering her as a fairly admirable person. By only claiming that character traits are tendencies, the situationist may therefore be able to capture all the cases she has in mind.

Now, the problem for the situationist is that, the more she allows some degree of independence between character traits and their manifestations, the less strong her case becomes. As we saw, the situationist would have to provide further reasons for her interpretation of the data if she were to accept the blunt claim that character traits can exist without ever manifesting themselves. This is why, we saw, the situationist may argue that it is part of the very idea of being a trait of character that it be reliably connected to its manifestations—more precisely, that the existence of the former depends on the occurrence of the latter—giving the situationist the right to infer from the data the absence of the relevant traits. But is it really true that it is conceptually impossible for a character trait to exist without manifesting itself in behaviour? Even if it is true

\(^{79}\)See also Adams’ (2006, 124) distinction between virtues related to perfect duties and virtues related to imperfect duties.
of some, even all, virtues, why think that it must be true of character traits in general?

Consider the following case. John is a young soldier who has proved to everyone that he is a profoundly courageous man. While in war, he has been courageous not only on the battlefield, both in fighting the enemy and in helping his friends, but also in the way he could stand up against his superiors when the latter asked him to perform what he takes to be horrendous and unfair acts (e.g., killing innocent civilians). After being expelled from the army for disobedience, John comes back home and, upon realising that the war was a mistake, becomes depressed for an extended period of time, and, as a result, behaves in ways that are radically different from before. Above all, he fails to be courageous when the situation calls for this quality. It is fair to assume that, were the depression to go, John’s courage would manifest itself again. But it is also fair to say that John, while depressed, is not very courageous. Yet, it is not impossible that the very trait that manifests itself when John performs a courageous act remained intact, simply being masked by the depression. This underlying trait, we said, is not a virtue, as (we have assumed) a trait cannot be a virtue if it does not reliably manifest itself. However, it doesn’t seem crazy to say that it is part of who John really is, part of his character. So, even though virtues may be conceptually required to manifest themselves, no such restriction seems to be in place with respect to character traits as a class. By introducing the generic concepts of character and character trait, therefore, into the list of concepts under dispute, we can see that the situationist may be guilty of some sort of selection bias.\textsuperscript{80}

If the situationist were to accept the tendential view of character traits, therefore, she would have to tell us what the dormant disposition is in cases such as the one just given. In particular, she must say why the disposition

\textsuperscript{80} It is important to note that, at least with some level of idealisation, there is a reliable connection between the dispositions that make up character and their manifestations. The point of the discussion is only that, in less than ideal conditions, such a connection may repeatedly fail to be made such that the relevant dispositions can exist without their being manifested.
cannot count as a character trait. So now we can complete the problem posed in the previous section. We have seen that, if the situationist were to accept the distinctness of character traits and their manifestations, she would need an independent argument in order to show that people are influenced by situational factors because they don’t have the relevant traits (rather than because their traits are masked, say). The argument was that, given that the concepts of character should be interpreted tendentially, we can use the absence of apparent manifestation as evidence against the presence of character traits. My reply is simply that not all of the relevant concepts should be interpreted tendentially, or at least that it is not obvious that they should be interpreted this way: there is a further reasonable debate to be had on the matter. Until the situationist provides an argument that does not overly rely on the specific case of the virtues, the ontological claim according to which character traits are non-existent or rare cannot be defended on the basis of the data on which the situationist relies.

In the next section, I will provide the outline of an account of character traits, and their relationship to virtues, that will give the virtue ethicist the tools to respond to the situationist challenge by providing an alternative interpretation of the data: that seemingly trivial situational factors can prevent people from manifesting their character, and therefore from being the virtuous people that they could be.  

6. A non-reductive, non-tendential account of character traits

We have seen that both the reductive and tendential views of character traits are inadequate in accommodating independently plausible theses. Against the reductive view, we have seen that it fails to accommodate the idea that character traits are productive of mental states and behaviour. Against the tendential view,

---

81 Or vicious. It goes both ways: situational factors may prevent people’s bad character traits from turning into outright vices (assuming here that vices must reliably manifest themselves, which may not be as obvious as in virtues).

82 Nothing I say here precludes us from thinking that the fact that people’s character can be masked as easily constitutes a significant moral fault, or at any rate that it is reason for pessimism. The focus of the discussion here is however on, so to speak, the descriptive side of character and virtue concepts.
we have seen that it fails to capture the conceptual possibility of character traits without tendencies, or at any rate fails to provide an explanation of why dormant dispositions—things which it readily allows—cannot count as traits of character. A generic notion of character, therefore, seems missing from the situationist’s project of undermining virtue ethics. However, it is not clear that the situationist needs to accommodate it, and not clear that she must have an account of character traits that captures them. After all, if virtues are tendencies, then the situationist can straightforwardly rely on the data in order to claim that virtues are at best rare. Things are not as simple, however, or so I would like to suggest.

First, let’s formulate a rough account of character traits that is neither reductive nor tendential. Presumably, it must claim that character traits are in principle things that can exist unmanifested. This would enable it to accommodate cases such as that of the soldier above. Several platitudes can then be shown to be compatible with the view; for instance, that character traits ‘define’ people (make them who they are), that they can contribute to people’s worth as persons, that they are related to what people care about, and so on. On the view under discussion, therefore, a view we can call ‘robust realism’ about character, character traits are dispositions to exhibit a range of behaviours but that do not all have to manifest themselves in order to exist.

Even though character traits generically construed may be rightly viewed in this way, we have seen that at least some virtues, as species of character traits, can plausibly be construed tendentially. Does this mean that situationism can be reframed as a challenge, not to the existence of character or character traits, but to the existence of these traits, i.e., the virtues? After all, if virtues are to be defined as tendencies, then it seems that people at best are rarely virtuous, at least if we trust the data. But this ignores a possible response from the robust realist: that although a person may lack a given virtue at a given time, she may nonetheless have a trait of character that, if manifested reliably, would constitute the relevant virtue, but that, because of some recurrent situational factors, fails to manifest in the relevant way. After all, if virtues are a species of character
traits, namely those (good) character traits that reliably manifest themselves, then it may take a minor change in the environment for the relevant character traits to become virtues. Perhaps, indeed, the good mood in the phone booth experiment is, despite its apparent triviality, all it takes for a character trait to be on its way to become a virtue. The fact that it figures in its current triggering conditions (i.e., in the class of factors that play a role in its triggering) is surely surprising, and may be cause for pessimism, but should not count as a reason to reject the claim alluded to at the beginning that if virtues are not widespread across the population at the moment, relatively minor changes in the world may lead to their acquisition. 83

The virtue ethicist, therefore, has the resources to provide the following alternative interpretation of the data where apparently no trait-relevant behaviour is performed: that the relevant character traits are not manifested, not because they don’t exist, but because there may be something in all the relevant situations that prevents them from manifesting themselves. Of course, this is assuming that character traits do exist, a claim for which no argument has been provided. However, notice that the situationist is in no better situation if she ends up accepting the account just provided of character and of its relationship to virtue. Given that the data do not rule out the hypothesis that the relevant situational factors work as enabling/disabling factors with respect to given character traits, a view that claims, on the basis of the data, not only that virtues are rare but also that they cannot be acquired over time in the manner suggested by the opponent, presupposes the absence of the underlying traits rather than argues for it. As a result, since the situationist would like to use the data in support of her claim, rather than merely show that her claim can lead to a plausible interpretation of the data, she must provide an argument to the effect

83 I agree though with Christian Miller (2009a, 269) when he claims that a trait that requires the presence of good moods or empathetic feelings for it to manifest itself hardly can count as a virtue such as compassion. A possibility Miller neglects however is the possibility, alluded to here, that although the relevant traits may not currently constitute virtues, it may be the case that, over time, they could become virtues, a possibility he seems doubtful about (270-271, though see Miller, 2009b, 169-170). His doubt, I think, is partly based on his different construal of the relevant traits as things that, even reliably manifested, would plausibly not count as virtues as the virtue ethicist thinks of them, as he seems to think that at least many of their instances cannot be triggered without the presence of the relevant affective states.
that the robust realist account of character, and its relationship with virtues, is mistaken. Given that common-sense currently accepts the existence (and plenitude) of character traits, it seems that the burden of proof has shifted back to the situationist to provide independent reasons why this commitment should not be trusted.

7. Virtue ethics and weak situationism

Even though the burden of proof on the virtue ethicist may be discharged, and that the situationist data can be interpreted in character realist terms, it may seem that positive reasons to believe in character traits should nonetheless be provided. After all, as we have assumed, the data readily support the adoption of some form of scepticism or other about our ability to know about the distribution of character traits in the world, assuming they exist—what I called ‘weak situationism’. One thing to notice is that the realist about character traits at issue can readily accept weak situationism, for after all she is committed to the view that the character traits of other people can be systematically hidden from view. Yet, it seems that she shouldn’t be a sceptic about our ability to have knowledge about character traits full stop, for it would then be reasonable to ask how she knows that character traits exist. There must be, the worry goes, some way of getting in touch with other people’s character that is less epistemically problematic than the way we put to use in our day-to-day encounters with strangers (not to mention in our observations of subjects in social psychological experiments).

I contend that there are two main ways we can get at least justified belief about other people’s particular traits of character, one intellectual, the other more affective. The first way is roughly by observing individual people over time, meticulously noticing the situational factors that seem to play a significant contribution to their behaviour, and trying to find out whether the absence of such situational factors would render the relevant piece of behaviour less (or more) likely to happen. In addition, we can have a story as to the role particular situational factors play in any given situation, in particular whether the role they
play is typically that of an enabler or that of a disabler. Good mood, for instance, has been claimed to widen our attention in such a way that we are more likely to notice things that we wouldn’t have noticed in a neutral or bad mood (Fredrickson, 2003), suggesting that it may sometimes play the role of a facilitator (or a temporary mental scaffolding, so to speak) in the triggering of certain dispositions.\(^{84}\) Having such an account in turn would allow us to identify the kinds of traits that are likely to play a role in a given person’s behaviour. Of course, this would not warrant, on the basis of a single action, an attribution of vice to someone who, say, after having found a dime in a phone booth, nonetheless doesn’t help—after all, the person may take the dime to be of negligible worth, may be depressed, and so on. However, if we learn that this person rarely helps others \textit{even} when she is in a good mood, is not depressed, or is otherwise in very disposition-friendly environments, this warrants the claim that she lacks not only the virtue of compassion, but the character trait underlying this virtue as well. The fact that character traits can remain dormant does not mean that we should attribute compassion (or the trait underlying it) to someone who never helps others, especially when the conditions are friendly to it.

Even though we may not be justified in attributing particular character traits on the basis of scarce data,\(^{85}\) we can still be justified in thinking that people have character traits that, if we look closely enough, and if we spend enough time with them, we would be able to observe, at least in principle. I would like to suggest that this is what sometimes happens when the people in question are

---

\(^{84}\) There are surely alternative hypotheses about the role of good mood in (e.g.) helping behaviour. One is that it plays a more direct role in the determination of behaviour, in for instance making us help others in order to maintain our good mood, as claimed by the proponents the ‘mood maintenance hypothesis’ in social psychology (e.g., Isen and Simmonds, 1978, Forest et al., 1979). The point here however is merely to give an example of a situational factor that may work as a facilitator in the triggering of certain character traits—making us, for instance, more attentive to the world—and hence that would not cast doubt on the possibility that these traits become genuine virtues (as the relevant situational factor is not what ultimately motivates the behaviour). I suspect the view is rather close to the ‘concomitance hypothesis’ in social psychology (e.g., Batson et al., 1979). I suspect, however, that the interpretation of the data found in the social psychological literature may vary depending on one’s prior conception of moods (including the function moods are thought to have), and therefore that it is crucial to agree on what kind of state we should take a mood to be. See Miller (2009b) for a discussion.

\(^{85}\) Though see Bates (ms.) for confidence that the data can justify us in attributing certain character traits, namely \textit{vices}. 

people we *love*. As some have argued (e.g., Jollimore, 2011, Little, 1995), loving someone—or more generally caring about someone, as in friendship—implies being disposed to attend, often via emotion, to that person more than we would if she were not loved or if she were a stranger, and therefore to learn important facts about her. In addition to the natural attention that we give them, the sheer number of hours we spend with the people we love makes it significantly easier to observe them in various situations. Of course, the situationist may reply that any regularity in behaviour may be in fact more the product of stable situational factors (including the fact that the friend or lover is present) than the product of robust traits. At this stage, however, the point is not to show that this alternative hypothesis is false, but to show that the character realist can provide an epistemology that does not rely on the problematic capacities at play when we attempt to attribute traits of character to strangers.

Notice now that it is one thing to claim that we are bad at knowing other people’s traits of character, quite another thing to claim that for that reason we shouldn’t believe in character traits. Even though the epistemology just outlined may turn out to be mistaken (though see Ch.6), there may still be reasons to think that people’s psychology is structured in such a way that there must be persisting traits that dispose them towards certain behaviours in ways that character traits do. For instance, given a view of the world as composed of persisting causal powers that make physical objects behave in various kinds of ways (see, e.g., Molnar, 2003), there may be no good reason to deny that psychology is made up of the same building blocks, and that among them are dispositions that fit the description one might have of character. More crucially, given a view of the world as comprising *lots* of causal powers, there may be no good reason to deny that psychology is composed of very many traits that dispose people towards behaviours that are very often incompatible, making it unsurprising that many of them will find themselves recurrently masked. The point is not that a character realist such as the virtue ethicist must accept this ontological picture of the world; the point is rather that realism about character can be defended on grounds that are independent from the debate over

---

situationism. It is then up to the situationist to tell us what her independent grounds for scepticism are, requiring her, I suspect, to put the data to one side and start to engage in a more thoroughly theoretical inquiry about character instead.

8. Conclusion

In this chapter, I have attempted to put pressure on the situationist to provide an account of character traits as friendly to empirical testing, such that the social psychological data available can be used in support of her scepticism. I have argued that the two most plausible versions of such an account either are inconsistent with commitments the situationist should independently accept or are not general enough to include certain traits of character. Accepting, for the sake of argument, the situationist’s account of virtue, I have argued that, were the situationist to accept a plausible general account of character, her reliance on the social psychological data in support of her claim would lose its force. Worse, the account of character provided gives the virtue ethicist the tools to show the situationist that, even though virtues may be at best rare, there is at this stage good reason to think that, with the proper training (including, perhaps, training in attention) and some environmental aid, people may, over time, become more virtuous. This is, of course, assuming that people do possess character traits—both good and bad—a claim for which I have provided some tentative support. Overall, it may be premature to reject the virtue ethicist’s optimism that virtue is possible. We may have the tools to become virtuous; perhaps we should give ourselves a chance.
Chapter Six
Sentiments and Reasons

1. Introduction

Now that I have defended the reality of sentiments, I would like to investigate the question whether certain kinds of sentiments can be said to be valuable in some way. Of course, it is clear that not all kinds of sentiments as I define them will turn out to contribute positively to our lives—phobias, for instance, or ethnic hatred are things we agree that we should get rid of (if we can). And although such sentiments may be useful on certain occasions—fear and anger can help us face a difficult situation—it is rather uncontroversial to say that we would be better off without them. The sentiments that we should give up seem to be either morally faulty (ethnic hatred is morally wrong) or based on epistemically faulty mental states (the evidence does not warrant relevant beliefs). These considerations suggest that, while sentiments can be attitudes taken towards the world, they are not necessarily supported by adequate reasons. Perhaps indeed they are never supported by reasons. This view is too strong, as there seem to be at least moral reasons, and perhaps prudential reasons, to love rather than hate people from other cultures—morality may indeed require us to minimally care about the welfare of everyone, regardless of where they come from. Prima facie, therefore, there are reasons for sentiments. However, can there be epistemic reasons for sentiments? Can there be reasons that bear on the correctness or accuracy of the attitudes that constitute them?  

---

87 In the following, I’ll use the term ‘epistemic’ in a broad sense as concerned with the kind of justification possessed by mental states that both have accuracy conditions and are, in virtue of having accuracy conditions, subject to standards of justification (what I will call ‘cognitive’ states throughout this chapter). Such states include beliefs and emotions, but not perceptions (perceptions have accuracy conditions but are not subject to standards of justification). The belief that \( p \) is justified just in case there are reasons supporting \( p \), and an emotion of sadness is justified just in case (say) there are good reasons to take the object of the emotion as involving a significant loss for the subject. The question will be whether love and caring are subject to
The answer depends on what kind of attitude sentiments could be, and whether they can be correct or accurate in the relevant way. If all sentiments are mere conative states conceived as brute causal forces that trigger action without attributing any genuine property to the aimed at state of affairs (as some conceptions take desire to be), or if they are cognitive states that arbitrarily project value onto the world (by attributing to things in the world value properties that they in reality don’t have), then all sentiments would be either non-rational or irrational. Either way they would not be epistemically valuable, whatever other value they may have.

Assuming that some sentiments may be mere conative states, and putting aside the possible cases of sentiments that are not intentional mental states (or what I called ‘attitudes’), we can ask whether some sentiments not only include a cognitive dimension that is in principle responsive to reasons, but also are sometimes responsive to reasons. Are there sentiments it is epistemically rational to have? On the face of it, the answer seems to be yes. We do sometimes wonder why others love their partner, or what it is about their friends that make them suitable for their love. We also criticise the love of others by saying that their beloved is unworthy of their love. Moreover, caring in general seems to admit of cases where an instance of care is misplaced due to epistemic error. I can care a lot about my best friend until the day I find out that she’s been deceiving me the whole time into thinking that she cared about me; in such a case, it seems that I am provided reasons to stop caring about her. Love and caring, as a result, are good candidates for the status of rational responses. However, as we will later see, alternative accounts construing love and caring as either a-rational or irrational responses to the world may nonetheless have the resources to account for our practices of evaluation.

---

standards of justification (and sometimes satisfy them) in an analogous way beliefs and emotions are. I take it that if they are sometimes justified in this way, this would thereby constitute one way they are valuable to us. I remain however open to the possibility that there are other, non-epistemic, ways love and caring might be valuable. For instance, if all forms of caring turn out to be non-rational or irrational (in the relevant epistemic sense), they may still be necessary ingredients for a good life.

88 For instance Frankfurt’s view, described later in this chapter.
89 Of course, some may be instrumentally valuable in helping us acquire true beliefs.
In this chapter, I will argue that, although some forms of sentiment (including some forms of caring and love) are a-rational or irrational, some other forms can plausibly count as rational. After introducing cases that ‘antirationalism’ (namely the view that all sentiments are not rational responses to their object) can easily account for, I will tackle the question whether it can also accommodate one of the paradigmatic forms of caring, namely caring for something as the very thing that it is (as opposed to as satisfying a certain description). We will see that the ‘rationalist’ alternative, although supported by certain intuitions, suffers from problems, in particular in its attempt to capture romantic love and love between friends. I will end this chapter by assessing various rationalist solutions to these problems, and will propose what I take to be the most plausible.

I’ll mostly ignore metaethical considerations here, and take at face value our practices of evaluation of caring and love. In particular, I will ignore views that take such states to involve (at least in part) a representational content (to be beliefs or belief-like) but that claim that nothing in the world actually corresponds to such content. Even if such error theory is right, we could still ask about the standards of justification that we hold regarding our evaluative attitudes. We could still ask what makes a moral belief a justified one. For instance, even if moral error theory is true, we can still justify the claim that an act is wrong by appealing to the level of suffering that it would produce.

2. Rationalism vs. antirationalism about caring and love

Recall the claim made in Chapter 4 that a disposition is psychological if it is an attitude of some sort. We have defined an attitude as a mental state that is directed towards an object and presents it as being a certain way. On such a construal, we have seen, caring and love obviously count as attitudes. First, they are directed towards entities of the world (e.g., my love for my parents). And second, they construe their object in a certain way, under a certain aspect (e.g.,
my love construes my parents as the persons they are, and not as whatever aggregate of atoms that constitutes them).

Broadly construed, attitudes can be of various types, including beliefs, desires, imaginings, memories and so on, and their content, I have assumed, admits of propositional (e.g., belief and desire, construed standardly) and non-propositional forms (e.g., perceptual states). I have taken the attitude of taking as valuable as a placeholder for whatever attitude caring is. The term was intended to remain neutral on caring being a cognitive or a conative state.

It is important to note however that on appearance the phrasing ‘taking as valuable’ suggests a cognitivist reading. If one takes something to be valuable, it seems, one attributes a value property to it that it may or may not have. One therefore can be in error about the value of the object of one’s care. Another way to construe the notion of taking as valuable is as a state that, although construing its object in evaluative terms, doesn’t purport to represent the world as it really is—to use a overused metaphor, it doesn’t have a mind-to-world direction of fit. It is rather to be analysed as a purely subjective construal of it that is not subject to epistemological standards of any sort (analogous to the epistemological standards found in the case of belief). As Hume claims, “'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me.” (Hume, 1955, 458) The so-called ‘pro-attitudes’, with desire as paradigm, are often understood in this non-cognitive way (e.g., Smith, 1994, who qualifies Hume’s account of belief and desire as “the standard picture of human psychology”, 7), and Harry Frankfurt, when defending his antirationalist view (see below), seems to follow Hume in thinking that they are not subject to any rational appraisal (or at any rate to the rational appraisal that cognitive states such as beliefs are subject to).90

90 Frankfurt’s desires, at least those not derived from other desires one has (or ‘basic’), would in fact be on a par with mere urges (which can also be basic) (Kolodny, 2003, 144-145). One thing to note however is that the Humean can allow cases of irrational desires, namely those formed on the basis of irrational beliefs: “desires are subject to rational criticism, but only insofar as
It is customary to take caring and love as pro-attitudes of some sort, suggesting that, although they may generate reasons for action, they are themselves not supported by epistemic reasons. Construing such states in terms of (Hume’s) desires (or, if non-propositional, in terms of desire analogues) does not do justice to our ordinary understanding of them, however. If I love Sophie, I do not simply desire certain things to be true of her. On top of that, I take Sophie to be a certain way; I take her, say, to be the wonderful person that she is. This is the whole point of coining the attitude ‘taking as valuable’. In order to avoid confusion, I will therefore call ‘non-cognitive’ any state, chiefly the ‘pro-attitudes’ construed in Humean terms, that is not subject to a norm of representational accuracy—to be contrasted with ‘cognitive’ states, which are subject to such a norm.

The question with which I will concern myself in this chapter is whether caring and love, construed as attitudes that take their object as valuable, should be conceived as cognitive or non-cognitive. Let’s call rationalism any view that takes caring and love to be partly a cognitive matter, namely a matter of representing a world that is independent of them, and that claims that many of their instances are supported by reasons in favour of the representation they involve:

---

they are based on beliefs that are subject to rational criticism. Desires that are not related in some such way to beliefs that can be rationally criticized are not subject to rational criticism at all.” (Smith, 1994, 8) Of course, Frankfurt could claim that all forms of love and caring are constituted by basic desires, namely those not dependent on other mental states in this way. But I take it that if he wanted to save his view from the forthcoming argument to the effect that certain forms of love and caring are cognitive, Frankfurt would have to allow irrational instances of love and caring, namely those based on false beliefs. See Goldman (2009) for a view of desire as both cognitive and bestowing value as opposed to responding to it (à la Frankfurt). I will remain silent on the question whether this reply would be successful, in part because I’m not sure how to make sense of the combination of a cognitivist view of desire and a subjectivism about value. Furthermore, I am suspicious that the sense in which desires can be unreasonable on the Humean account is the same as the sense in which one’s love for (say) an abusing husband (see below) is unreasonable. In any case, further work needs to be done by the Humean who reduces caring and love to desire.

91 It should be kept in mind however that such a division is very crude, and that many states under the banner ‘pro-attitudes’ may include a cognitive element. Some (Oddie, 2005) take desires themselves to be cognitive in this way. For the sake of simplicity, I’ll take such states, if there are any, as cognitive.

92 It is important to note that rationalism here need not deny that caring and love have a non-cognitive dimension. It rather claims that, in addition to any non-cognitive dimension they may have, caring and love involve a content for which we can supply reasons.
Rationalism: Caring and love are cognitive states construing their object as valuable in a way that is independent of them, and can be adequately supported by reasons.

Antirationalism, by contrast, is a view that denies one or both of the two conjuncts accepted by the rationalist. It comes in two main forms:

A-rationalism: Caring and love are fully non-cognitive states construing their object as valuable without being subject to the epistemic standards applicable to cognitive states. Let’s call ‘attachment’ any such ‘a-rational’ state.

Irrationalism: Caring and love are cognitive states construing their object as valuable in a way that is independent of them, are in principle states for which we can be epistemically justified, but, contra the rationalist, are in fact states for which we are never epistemically justified. Let’s call ‘projection’ any such ‘irrational’ state.

The three views as stated are global in that they purport to account for all possible cases of caring and love. It is crucial however to note that they can come in both global and local varieties. One can be a local rationalist about a given class of cases and a local antirationalist about another class. (In the following, I’ll take for the sake of simplicity of exposition global antirationalism to be true just in case all types of caring and love are a-rational, or all types of caring and love are irrational, or some are a-rational while others are irrational.)

Given the variety of things that the terms ‘caring’ and ‘love’ apply to, it would be surprising if global rationalism were true. As we will see in the next section,

---

93 It is important to emphasise that the irrationalist need not claim that caring and love are necessarily irrational, that is, constitute projections in all possible worlds. What the irrationalist typically claims is that, as a matter of contingent psychological fact, caring and love make us attribute to their object excellences and perfections that are in fact not there, or at any rate that a suitable outside observer would have no reason to attribute to it.
there seems to be such a thing as ‘blind’ love. Furthermore, it is difficult to take a mother’s love for her child (or the child for her mother) as a matter of justified response to the child’s (the mother’s) valuable properties. The rationalist may reply that it depends on how we individuate types of cares and love. Perhaps cases of blind love and infant attachment are non-standard (some may say ‘perverse’) instances of the general type of love, which is governed by epistemic standards. The rationalist, after all, need not deny that caring and love can fail to satisfy epistemic standards. She could as a result say that the relevant cases involve irrational instances of love, instances that may derive their value (especially in cases of mother-child attachment) from non-epistemic considerations.

I can’t help finding this kind of move desperate, at least as far as mother-child attachment is concerned. First, there seems to be no non-question-begging reason to privilege the rationalist’s way of classifying forms of caring and love. Why should we think that instances of love for human infants to fall under a general kind of love that is (for the rationalist) subject to epistemic standards? Second, it doesn’t seem right to think of love for our children to be construed as irrational. As Frankfurt (2004) argues, such love doesn’t appear to be the kind of thing for which reasons could be appropriately given.

Of course, the rationalist could argue that, although instances of blind love are irrational, instances of love for our children are rational, perhaps even always rational. As we will see in the next section, there are strong reasons not to think of this kind of love to be subject to epistemic standards in this way. For now, it is better to follow common-sense classifications and accept the possibility that some types of caring and love are as the antirationalist claims, which means accepting that certain forms of care and love are either non-cognitive or systematically irrational. The question will be whether antirationalism can be generalised to all cases, i.e. whether global antirationalism is true.
3. Prima facie reasons to be at least local antirationalists

In this section, I will introduce cases that the antirationalist is particularly well-positioned to account for.

3.1 Projection

In the previous section, I gave the example of blind love in order to illustrate the possibility of forms of love that are, on the face of it, systematically irrational. Blind love occurs whenever a lover attributes excellent qualities to a person who lacks them, or at any rate who does not display adequate evidence for their presence. Such qualities are merely imagined being there in a way that gives the lover the impression that she is responding to independent value. As Stendhal famously remarks:

If you are sure that a woman loves you, it is a pleasure to endow her with a thousand perfections and to count your blessings with infinite satisfaction. In the end you overrate wildly, and regard as something fallen from Heaven, unknown as yet, but certain to be yours. (Stendhal, 2004, 75, quoted in Jollimore, 2011, 7)

Cases of irrational projection are pervasive in caring, not all of which can be as destructive as blind love. I can care about wearing particular boxers (my ‘lucky’ boxers) every time I have an interview. You can ask me why it is so important for me to wear them, what it is about the boxers that make me think that they bring me luck and as a result are worthy of care. My answer could be “Nothing, I just like the idea of having a little superstition in my life.” In such a case, I am well aware that my care is unwarranted, and yet wearing them makes me confident and hopeful, and failing to wear them makes me frustrated and angry. In short, the irrationalist seems to be right to emphasise our ability to attach to things in a way that is almost arbitrary, as well as our laxity when it comes to the critical assessment of our caring attitudes in light of the evidence we have.
Cases of irrational cares give us a reason not to think that all instances of caring and love are rational. This, however, is something that the rationalist can accept. The rationalist makes the claim that love is a cognitive state that is in many cases epistemically justified. She can therefore agree with the irrationalist that cases such as blind love involve unjustified instances of love, thereby counting as a ‘local’ antirationalist. The rationalist, however, has the burden of giving us an account of how caring and love can be epistemically justified when so many instances of them seem in deep conflict with the relevant norms.

3.2 Attachment

Prima facie, therefore, there are good examples of cognitive forms of caring and love, even though such forms may only have irrational instances. This by itself should already give us a reason not to adopt a global form of a-rationalism. Blind love, it seems, involves an epistemic mistake. But appearances may be deceptive. Perhaps, the a-rationalist could say, blind love is unfortunate, not because it is unsupported by adequate reasons, but because it is potentially destructive, or because it is often unrequited, or because of some other, non-epistemological reason. Perhaps, moreover, it is crucial that the lover have adequate beliefs about the beloved in order for his love to be directed at her; a love that fails to pick out the relevant person, and is directed instead at an imagined one, is something we don’t really want.

Having said that, the a-rationalist does not typically appeal to cases such as blind love in order to argue for her position. The paradigm of a non-cognitive form of love, rather, is the love of a parent for his or her child. As a major proponent of a-rationalism, Harry Frankfurt writes:

[L]ove does not require a response by the lover to any real or imagined value in what he loves. Parents do not ordinarily love their children so much, for example, because they possess exceptional value. In fact, it is the other way around: the children seem to the parents to be valuable, and they are
valuable to the parents, only because the parents love them. Parents have been known to love—quite genuinely—children that they themselves recognize as lacking any particular inherent merit. (2006, 25)

For Frankfurt, the kind of love that parents have for their children is subject to no norm that requires them to first identify valuable features in the children that would warrant their love. It would preposterous, it seems, to ask all parents not to love their children until the latter have proved themselves by acquiring valuable properties. The common case of loving our children while they are still in the womb would turn out to be a common mistake. As Frankfurt notes, the fact that a parent’s love for his child is not a response to the latter’s valuable features does not imply that the parent takes the child to be valueless. On the contrary, by loving her, the parent makes the child valuable in a way that she wasn’t before. 94 As is sometimes said, the parent bestows value on the child. Love is not responsive to reasons, on this account, not because it is irrational, but because it is non-cognitive. There is no sense in which we can be mistaken in what we are attached to, on this story. And although the attachment may be based on false beliefs about its object, it cannot be criticised on this basis, the same way the urge to eat caused by a mirage is not by itself subject to epistemic criticism. 95

Frankfurt’s a-rationalist account is very plausible when it comes to capturing cases such as a parent’s love for his child. We clearly wouldn’t criticise a parent’s love for his child if the child turned out to be a moral monster. However, Frankfurt goes further in claiming that love in general—and in particular love between persons—is non-cognitive as well. As we have seen, the a-rationalist can account for cases of apparent irrational care and love. Can he

---

94 The value being acquired by the child being presumably relational. Frankfurt does not give a clear account of how such creation of value is possible. It is plausible however that he subscribes to a subjectivist account of value according to which to be valuable just is to valued, at least when it comes to attitudes such as caring and love. See Frankfurt, 2004.

95 After all, an outside observer wouldn’t say here “You shouldn’t be hungry; there is no food over there”, as opposed, perhaps, to “You shouldn’t think that there is food over there; it will make you hungry”.

170
similarly account for instances of caring and love that are on the face of it neither a-rational nor irrational? After introducing in the next section considerations in favour of a negative answer, I will assess arguments to the effect that love for particular people should be conceived along a-rationalist lines.

3.3 Does the account generalise? Two observations

Can Frankurt’s a-rationalist view be generalised to all cases of caring and love? There are two considerations that should make us find this possibility suspect. First, it seems to be a datum of common-sense that some types of things are more worthy of our caring attitudes than others. There are certainly things it makes more sense to care about or love. For instance, someone who loves phonebooks more than she loves humans seems to be making, not merely a prudential or moral mistake, but an epistemic one. Why? Because humans in general have properties that make them suitable for caring in a way that phonebooks in general aren’t. There is barely anything about phonebooks that could justify our deeply caring about them. To be sure, a phonebook in particular could be important to us because of its connection with other things we care about, such as the fact that it belonged to someone we loved or that it comes from the phonebook factory where we worked before it closed. It is clear however that, however appropriate it is to love phonebooks in general, it is not appropriate to love them more than people.

The first observation claims that there are certain types of things that are more worthy of our love than others. A second observation is that tokens of the same type can be more worthy of our love than others. There are instances of things of particular types that it seems we should care more about than others. Some people, for instance, seem to be more worthy of our love than others. To take a seemingly uncontroversial case, one’s love for Mother Theresa is appropriate in a way that one’s love for Ted Bundy isn’t.
In both cases, entities in the world (i.e., the various properties that objects and persons instantiate) seem to provide us with reasons to selectively care about or love certain things at the expense of others, reasons that are independent of the attitude.\textsuperscript{96} The fact that I mostly love people, and not phonebooks and rolls of toilet papers, is not accidental (Jollimore, 2011). When I love something or someone, it seems, there is something that I find valuable about it or her, and something that I can be wrong about if it turned out that the entity lacked it. Not everything is worthy of our love. Let’s call the intuition according to which some things, in virtue of their properties, merit our love whereas others, with different properties, don’t the merit intuition.

There is ample evidence that the merit intuition is at play in much of our thinking on caring and love. We do often ask ourselves if a given activity is worth our while, and whether there are alternative options that may be more valuable. According to the rationalist, asking these kinds of question is a matter of asking whether caring would be an appropriate attitude to have towards the relevant activities. And this seems to be right: deliberation about what’s worth caring about is something we surely do. Another piece of evidence is our judgements to the effect that we care too much or too little. If we take these judgements at face value, there are facts about objects that are relevant to, not only whether we should care about them, but also how much we should care about them. Perhaps indeed we should in the end care about phonebooks (given some history they may have), but as we have seen this does not mean that we should care about them as much as we care about other things (e.g., other humans). Caring is clearly not, or not always, the non-rational attitude that the a-rationalist has been talking about.

How can the a-rationalist respond to these considerations? There are at least three options open to the a-rationalist. First, she could challenge the way I have described the intuition at play in our thinking about caring and love. She could

\textsuperscript{96} Of course, such reasons may in some way still depend on our psychology or our constitution for their existence. As throughout this chapter, I would like to remain neutral on such metanormative issue. The crucial point is that some instances of caring and love appear to be supported by properties that are independent of them.
for instance claim that, whenever we say that someone or something is unworthy of our love, we actually mean, say, that “the cost of loving them would be greater than the benefit of doing so; or perhaps it means that to love those things would be in some way demeaning.” (Frankfurt, 2004, 39) It may be the case that our judgements of worthiness are not always judgements about the appropriateness of caring or loving on account of the properties of the object. Perhaps it would make sense for Romeo to say that Juliet is not worth it, meaning that a union with her would make his life hellish. In addition, it may make sense for a noble woman to say of a working-class man whom she is attracted to that falling in love with him is not worth the cost of shame that would ensue. Note however that in both of these cases, we are not in the least tempted to paraphrase the claim by saying that the person in question is not worthy of our love. Rather, we make a claim about the circumstances that would ensue if one loved the other. If the two claims are indeed distinct, because they are in fact about different things, then the a-rationalist cannot successfully reduce one to the other.

The second option is to grant that the intuition should be understood along rationalist lines but then claim that it is mistaken. According to this response, it is simply not true that certain things are worthy of our love whereas others are not, or that things can be more or less worthy of our love. How can the a-rationalist motivate this view? She could simply argue that since there is no such thing as a value that is independent of our caring attitudes, we are mistaken in thinking that certain things are worthy of our love in virtue of their value properties. Although there is a lot to be said about the theory of value on which such a response relies—after all, it may be true that value properties are, e.g., mere projections of our sentiments—this option fails to appreciate the fact that, even if objective value properties turned out not to be part of the fabric of the world, this wouldn’t show it isn’t part of our perspective as valuers that our evaluative attitudes can be more or less justified, more or less appropriate, with respect to their object. And it is from this perspective that the question whether caring and love can be justified should be asked. Moreover, even if the strategy of showing that the intuition is mistaken by defending an antirealistic account of
value were successful, this wouldn’t give us reasons to favour the a-rationalist version of antirationalism. For the mere fact that we have the merit intuition suggests that the forms of caring at issue purport to be cognitive. That values don’t enjoy an existence independently of our cares at best shows that irrationalism is right.

One last option for the a-rationalist is to concede that the examples of caring that the rationalist appeals to are indeed examples of cognitive states yet claim that they are not paradigmatic. As we have seen, humans are things it makes sense to care about. If a thing is human, then it probably has the features that make caring about it appropriate. The same may be true for specific humans. If an individual has an excellent moral character, for instance, then this may give us more reason to care about him or her than to care about an individual that is morally depraved. In our example above, it certainly makes more sense to care about Mother Theresa than to care about Ted Bundy. Note however the way the object enters into the content of the relevant cares. The rationalist holds both that we have reasons to care about whatever is human (or has human characteristics) and that we have reasons to care about whoever has an excellent moral character. In both of these cases, the particular identity of the individual who possesses the relevant properties appears irrelevant to our caring about her.

We care about the individual because she happens to satisfy conditions that could in principle be satisfied by someone else. To use some bit of philosophical jargon, the content of our love is thereby de dicto. However, the paradigmatic case of love for people is love that is directed at them as the very beings they are, and not as the kind of individuals they happen to be; paradigmatic love is de re (Kraut, 1987). It is thus essentially personal, to be distinguished from the impersonal stance that caring for humans in general exemplifies.

I find this response very compelling. It is true that the forms of caring and love that the rationalist has been appealing to are not (or need not be) the ‘deep’ forms of caring and love that we typically have in mind and for which I have given a partial account in Chapter 4 (where I defined caring as a ‘central’
disposition). So, even if the rationalist is right to claim that certain forms of caring and love are rational responses to the world, she hasn’t given us any reason to think that more paradigmatic forms of love (including presumably, for the a-rationalist, a mother’s love for her child) should be accounted for along rationalist lines. Progress has been made by the rationalist, however, as she has adequately shown that a-rationalism fails to generalise to all cases of caring and love. In the next section, I will discuss an argument for the claim that a-rationalism should be generalised to all forms of personal caring and love.

4. Love for particulars

‘Personal’ love seems to be the most paradigmatic form of love. Calling this form of love ‘personal’, however, does not do justice to the fact that we seem to be able to love ‘de re’ things that are not persons, animals being a clear example. So, because I don’t want to exclude outright such possible objects of love, we can call the relevant kind of love ‘love for particulars’ or, to use Kraut’s apt terminology, ‘de re love’. That being said, love for persons remains the most paradigmatic form of love for particulars, and the following will be mainly concerned with it.97

What reason is there to think that love for particulars is always a non-cognitive matter? On the face of it, the fact that we do not love the same people even though we can agree on what properties they possess suggests that the properties need not play any significant role in the formation of one’s love, and need not rationally compel one to start loving the relevant person. Frankfurt agrees: “As I am construing it, love is not necessarily a response grounded in awareness of the inherent value of its object. It may sometimes arise like that, but it need not do so.” (2004, 38) And even putting aside the case of a parent’s love for his child—Frankfurt’s favourite example—it is clear that many

---

97 There are arguments to the effect that persons have properties that make love (‘de re’) for them more appropriate than love for other animals. The property of being a rational creature may be one such (Velleman, 1999). However, there seems to be a tendency in the recent literature to think that loving animals in the way we love (say) our friends is not just less appropriate, but outright inappropriate (e.g., Helm, 2010). I would like to remain silent on this issue, although I’m inclined to think of such views as insufficiently motivated.
instances of love are not the product of rational inquiry into the nature of the beloved. Frankfurt continues:

Love may be brought about—in ways that are poorly understood—by a disparate variety of natural causes. It is entirely possible for a person to be caused to love something without noticing its value, or without being impressed by its value, or despite recognizing that there really is nothing especially valuable about it. It is even possible for a person to come to love something despite recognizing that its inherent nature is actually and utterly bad. That sort of love is doubtless a misfortune. Still, such things happen. (*ibid.*)

Given that value is not “an indispensable formative or grounding condition of the love” (*ibid.*), Frankfurt argues, rationalism should be rejected.

It is easy to see what the rationalist would think is wrong with such an argument. What is wrong with it is that it assumes a view of epistemic justification that takes mental states to be justifiable just in case they are the result of a rational process of inquiry. What Frankfurt overlooks here is the possibility that, although many instances of love are based on insufficient evidence—at least at their outset—there *may* still be evidence that could in principle justify them. Take the analogous case of belief. It seems that being formed on the basis of evidence is not necessary for belief; one can indeed be led to believe in a certain proposition “by a disparate variety of natural causes” (drugs, self-deception, wishful thinking, etc.). Does this show that belief in general is non-cognitive? Does this show, moreover, that it cannot be supported by adequate reasons? Surely not. The aetiology of particular instances of belief is irrelevant to the question whether belief in general can be supported by reasons. The same consideration should apply to love. The question we should therefore ask is whether love is the kind of thing for which epistemic justification is *available*, that is to say whether it is possible for a lover to be justified, regardless of *when* or even *whether* this justification is indeed
acquired. And Frankfurt so far hasn’t given us any reason to think that the answer is negative.

At this point, Frankfurt could appeal to his preferred case, namely a parent’s love for his child, in order to argue for an a-rationalism about *de re* love. For, as we have agreed, nothing I have said above precludes him from making this move. The question now is whether the considerations in favour of conceiving of a parent’s love for his child as non-cognitive are also considerations in favour of conceiving of, *inter alia*, romantic love and love between friends as non-cognitive as well. Recall the intuition that I have granted Frankfurt with respect to a parent’s love for his child, namely it does not make much sense to criticise one’s love for one’s child by citing some of his bad features. Nothing that the child could do, it seems, however atrocious it is, should give a mother a reason to give up her love for her child. To put the point vividly, it seems to be completely intelligible, and in no way rationally impermissible, for you to love Ted Bundy if you are his mother.

Of course, the rationalist could say that in such cases, the child *does* have properties that warrant loving him if you are his mother, properties that are independent from the bad ones just cited. And we’ll see in Section 7 an example of an account that allows this sort of response. But even if we were to grant that the parent’s love for his child is outside the purview of epistemic justification, we could still deny that similar considerations apply to cases of *de re* love between adults. For, as hinted at in the previous section, we *do* take particular people to be worthy or unworthy of our love. Take the oft-cited case of the abused wife who is given reason to think of her love as completely inappropriate given her husband’s clearly vicious character. She loves her husband, and she loves him as the particular person he is; yet, the fact that he mistreats her in the way he does clearly shows that she shouldn’t love him, that her love is not (or no longer) appropriate. The merit intuition at work again.

The reasons Frankfurt gives for taking familial love as non-cognitive therefore lose their appeal when it comes to love between previously unrelated people, as
in romantic love and love between friends. While it may be misguided to ask why a mother loves her child, it doesn’t seem so inappropriate to ask someone why he loves his partner. In addition, we do make judgements about how worthy of love someone is. For instance, we sometimes say of a friend that she needs someone who can love her as she really is (meaning, it seems: whose love is an appropriate response to her). To be sure, much of romantic love is not the product of the kind of careful inquiry that Frankfurt thinks the rationalist takes love to require, but this does not imply that it cannot ever be appropriate or inappropriate to love our partner or friend.

Now, the irrationalist could complain that the fact that the merit intuition is at play in our thinking about love at best shows that love is cognitive, not that love is ever appropriate. This complaint is fair, and in the following it will be important to keep it in mind. For we’ll have to have some account of justified love that does not ignore the fact that many instances of love are epistemically problematic. There are problems however that the rationalist about love should face in order to establish her position. We will see next that without an adequate rationalist response to these problems, the a-rationalist can declare her view superior.

5. Two puzzles about romantic love

What drives the merit intuition, it seems, is our unwillingness to accept the thought that our loves can be arbitrarily given to just anyone, irrespective of what they are like. This betrays an implicit adherence to a norm of appropriateness, allowing in turn the possibility of gaps between what our loves ‘say’ about the world and what the world is really like. The a-rationalist, by contrast, does not allow for such a gap. If talk of appropriateness were warranted here (perhaps in the minimal sense that there is nothing wrong with the love per se, perhaps in a similar way there might be nothing wrong with hunger per se), all instances of love would automatically count as appropriate. This can’t be right, however, given what has been said in the previous section.
At this stage, the a-rationalist has the right to ask the rationalist to give details as to how love can be epistemically justified. Since she has properly motivated her view, the next step for the rationalist is to develop it in a way that accommodates other desiderata we might have for an adequate theory of love. This, we will shortly see, is no easy affair.

Let’s start by proposing a natural way to construe rationalism about *de re* love (henceforth simply ‘love’). On what I will refer to as the ‘standard picture’, love involves the attribution of *intrinsic* value. Intrinsic value, moreover, supervenes on intrinsic properties. Finally, intrinsic properties of the object of love are supposed to render love appropriate or inappropriate.

The challenge for the rationalist is then to identify the kinds of properties in virtue of which love can be justified. A possible response is to say *any property in virtue of which the object is valuable*: being good-looking, smart, generous, creative, funny, and so on. The more one has any of these properties, the more it would be justified to love one. We can allow, furthermore, certain properties to have more weight than others in the justification of one’s love; perhaps, for instance, being smart contributes more to the justification of one’s love than being good-looking (or vice versa). Perhaps we can also allow personal tastes to play a role here. As long as there are intrinsic properties that, from one’s perspective, could (and sometimes do) render one’s love appropriate or inappropriate, rationalism is well supported. Although very attractive, however, the resulting account faces two problems, at least when applied to romantic love.

The first problem is the result of an implication of the rationalist position just sketched. On the standard picture, what make love for someone appropriate are intrinsic properties in virtue of which she is valuable. Assuming that many people have such properties, many people are worth loving. Does this imply that

---

98 ‘Taking as valuable’ being therefore construed as cognitive.
99 ‘Appropriate’ here can mean ‘fitting’ (or ‘correct’) or ‘justified’. In the following, I will understand the term as meaning ‘justified’, leaving aside the question whether love can be ‘correct’ in any robust sense.
we should love (i.e., *de re* love) *all* these people? Of course not; after all, we don’t have the cognitive resources to reach that state (probably only God could). And it is not because it is safe to assume that many people have the relevant properties that *we* are justified in loving them, for in order to be justified, it seems, we first need to be acquainted with them in some way. The rationalist need not even claim that the relevant properties, insofar as one is aware of them, rationally *oblige* one to have such an attitude towards anyone who has them. The rationalist could instead claim that the relevant properties make it rationally *permissible* to start to love anyone who has the relevant properties. Why love does not seem to be rationally obligatory in this way is an interesting question that I shall leave aside (although I’ll give the outline of an answer in Section 9).

But now rationalism seems to lead to an unacceptable conclusion. It seems to lead to the conclusion that in a case where two people are (say) happily married, and love each other for their valuable properties, it is rationally permissible for them to transfer their love to someone else just in case that person happens to possess the relevant valuable properties. In other words, rationalism leads to the thought that, were an individual sufficiently similar to, or perhaps even were an intrinsic duplicate of, my partner to come along, I would have as much epistemic reason to love her as I would have to love my partner; and given that the two would be very similar (perhaps *exactly* similar), there would be no epistemic reason not to substitute one for the other. A love for the duplicate would be as appropriate as my love for my partner. For many, this is an unacceptable consequence of rationalism. Let’s call it the *problem of substitutes*.

The problem of substitutes asks the rationalist to tell us why loving a substitute is not appropriate when whatever justifies the original love is also present in the substitute. Another problem for the rationalist is that loss of the properties that initially justified one’s love for another does not *ipso facto* justify the abandonment of the love, or at any rate make it rationally permissible. A love conditional upon the other keeping his good looks, intelligence or wit hardly looks genuine. So if the loss of such properties does not render one’s love
inappropriate, why should they render it appropriate in the first place? Call it the *problem of lost properties*.

Both problems cast doubt on the thought that love can be justified by the intrinsic properties of its object. While the former shows that the standard picture forces us to think of love as legitimately accepting substitutes, the latter shows that it commits us to the thought that one can appropriately cease to love another if she were to lose some of her valuable properties. Together, the problem of substitutes and the problem of lost properties suggest that love is an attitude of the wrong sort to be the subject of epistemic assessment.

Now, the a-rationalist can praise herself for avoiding both of these problems. Given that she takes love to be an affair of non-cognitive attachment, it is not surprising that love can remain untouched (and that we have the strong intuition that in some sense there is nothing wrong with it remaining untouched) whenever a substitute comes along or the beloved loses some of her attractive properties. So, however intuitive the merit intuition is, it may be a good idea not to take it too seriously given how counterintuitive are the thoughts that it leads to. The a-rationalist seems to have managed to show how unreliable the merit intuition is and thus that we shouldn’t give it the authority the rationalist grants it.

At any rate, this is right only if the rationalist cannot satisfactorily solve the problems. We will see in the next section that several options are open to the rationalist regarding the problem of substitutes. It should be noted before however that the force of the problem of lost properties, and the intuitions that it deploys, should not be overstated. Is it really clear that the loss of *all* the valuable properties that initially justified one’s love does not render it inappropriate? I think it is clear only if we make certain rationalist assumptions about the object of the love. Has she become a deeply depraved person without any redeeming qualities? If a caring husband were to become an abusive one, would that not make one’s love inappropriate? Loss of properties *does* sometimes justify the abandonment of one’s love. When it does, it may be
because the beloved has not only lost his valuable properties but also gained bad ones, or because the valuable properties that he lost used to have an important weight in the justification of one’s love (one’s love should perhaps survive the loss of good looks, but should it survive the loss of kindness?), and that no other weighty valuable properties have been gained in the meantime. What matters for love, it seems, is not whether the properties that initially justified love still exist, but whether some properties or other at any given time can make love appropriate.\textsuperscript{100} The rationalist thus can accommodate our intuitions about lost properties while continuing to accept the standard picture.

One may object though that even in the case where the beloved has lost all of her valuable properties, continuing to love her nonetheless is perfectly intelligible, a similar way perhaps Ted Bundy’s mother’s continuation of love for him is perfectly intelligible. Does this not suggest that there is at the very least something beyond the intrinsic properties of the beloved in virtue of which the continuation of love is intelligible? Could this sort of thing be a form of non-cognitive attachment? As I will suggest in the next section, we might have an inclination not to stop loving the one we love on account either of the fact that they have lost some of their properties or of the fact that substitutes are available. This does not yet mean that we may not also have good epistemic reasons to continue loving (as seen in the previous paragraph) or to refuse substitutes (see later sections). As long as love is not reduced to them, the existence of a-rational attachments is compatible with rationalism.

6. Options for the rationalist

We have introduced in the previous section two problems for the rationalist who subscribes to the standard picture. We have then seen that one of the problems—the problem of lost properties—can be solved within the confines of the view. Still, our reluctance to stop loving on account of a loss of properties may be explainable by appeal to an a-rational attachment. So, even if

\textsuperscript{100} Troy Jollimore (2011) proposes the interesting thought that past properties of the beloved can justify one’s current love. He says: “People are often loved for being beautiful, but it is also possible to love someone for having been beautiful.” (2011, 139)
rationalism is compatible with our intuitions about cases of lost properties, it cannot be established on this basis. In the following, we will mainly deal with the problem of substitutes.

The problem of substitutes cannot be solved the way the rationalist solved the problem of lost properties. Here, we cannot appeal to redeeming intrinsic properties, or to variations in the properties’ weight. In fact, we can hypothesise the object of one’s love as possessing many properties in virtue of which love for it is very appropriate. The problem is that, were an adequate substitute—with all the relevant valuable properties and perhaps some more—to come along, it would be on the rationalist view at the very least rationally permissible to transfer one’s love to that substitute. This, we have seen, may give us reason to think that love is a matter of a-rational attachment and not something that can be justified by appeal to intrinsic properties of its object. If the problem of substitutes were to force us to adopt a-rationalism, moreover, then the rationalist’s response to the problem of lost properties loses its force. An adequate response to the problem of substitutes is therefore needed for rationalism to be saved.

Fortunately, several options are open to the rationalist. At the most general level, we can distinguish two main strategies. The first strategy is simply to bite the bullet and discard the relevant intuition (that it would be inappropriate to accept the substitute) as unwarranted. Here’s a remark by Derek Parfit (1984, 295):

Suppose that I am a person who has moved into this different community. I fall in love with Mary Smith. How should I react after she has first used the Replicator? I believe both that I would and that I ought to love her Replica. This is not the 'ought' of morality. On the best conception of the best kind of love, I ought to love this individual. She is fully psychologically continuous with the Mary Smith I loved, and she has an exactly similar body. If I do not love Mary Smith's
Replica, this could only be for one of several bad reasons.

Of course, this sort of thought can be motivated by independent considerations about, for instance, the nature of personal identity. We need not go into the details of these considerations here. Another way to argue for this thought is by saying that what needs to be accounted for is not why it may be inappropriate to accept the substitute, but why it appears to us to be inappropriate. While there may be nothing wrong with accepting the substitute, the phenomenology is such that we take the people we love as irreplaceable (Badhwar, 2003).

The main problem with this kind of strategy is that it is insufficient to establish rationalism as better than a-rationalism. For the considerations that the rationalist appeals to may be compatible with a-rationalism as well. It is implausible that the a-rationalist must be committed to a particular view of personal identity—and for rationalists like Parfit, to a mistaken one—in order to hold her view. Furthermore, the a-rationalist well accepts the fact that the phenomenology of loving (or, on my account, of love’s manifestations) is such that it takes people to be irreplaceable. In fact, if anything, a-rationalism explains why things appear to us in this way: they have this appearance in virtue of the very kind of attitude that love is, namely a non-cognitive state that attaches itself to things without having to respond to their properties. What the bullet-biting rationalist needs to do, then, is appeal to independently plausible considerations that are incompatible with the truth of a-rationalism. She must, in other words, show why even the a-rationalist should give up on the thought that there is something fishy (although not ‘irrational’ in the relevant epistemic sense) about going for the substitute.

I leave the rationalist to develop this response. I think however that a better strategy for her would be to find an account of justified love that respects our intuitions about substitutes (and lost properties), even if this may mean giving up or modifying the standard picture. I take it that there are three main ways to pursue this second strategy. (1) We can show that in case of substitution, there
is a lack of epistemic justification for trading up, explaining our intuition in terms of an acceptance of an *epistemic norm*. Alternatively, (2) we can show that in such a case trading up may be epistemically justified but nonetheless in violation of a *non-epistemic norm*. Finally, (3) we can accept the rational permissibility of accepting the substitute but explain our reluctance to do so by appealing to the presence of an *a-rational inclination* that may, for some reason or other, be worth preserving (e.g., as just a harmless natural quirk of ours, possibly contributing to our happiness).

Note that solutions (2) and (3) are straightforwardly compatible with the ‘standard picture’ introduced earlier. They both claim that it would be epistemically appropriate to accept the substitute, even though other considerations may make it reasonable or rational not to do so. If we pursue (3), for instance, whatever reason the a-rationalist could give us for refusing to accept the substitute—for instance, that it would betray a lack of genuine love towards the original, or would be plain bizarre, or whatever—is in principle something the rationalist can accept. Love may indeed involve an element of non-cognitive attachment that inclines us not to accept substitutes. This, however, is compatible with the idea that substitution would nonetheless constitute the violation of an *epistemic* norm. In fact, holding the claim that love may sometimes or even always involve an element of non-cognitive attachment is compatible with *any* rationalist position that does not *reduce* love to a simple cognitive state. The challenge for the rationalist, therefore, is to motivate the thought that, whatever non-cognitive element love may involve, it is not *all there is* to it.

Notice further that a clear advantage of all these solutions over the bullet-biting one given above is that they all grant that there is something to be said for the intuition that love does not accept substitutes. Yet, only the first solution makes the bold claim that a love that accepts substitutes would constitute an *epistemic* failure (on top of other possible kinds of failure, such as moral and prudential). On the resulting account, there is something about the nature of the justification of one’s love for the original that does not allow it to transfer to a love for a
substitute. Given that this ‘epistemic’ strategy has been the one most pursued in the recent literature—perhaps because it would do full justice to the merit intuition or perhaps because it would make rationalism more elegant and simple than the alternatives—this is one the one that I shall pursue in the rest of this chapter.

It is still important to note that strategies (2) and (3) may deserve to be developed in greater detail. For instance, searching for possible norms at play in our intuitions about substitution is a worthy enterprise for the defender of (2). As for (3), it may be a good fall-back position for the rationalist in the grips of both the standard picture (and the merit intuition) and the intuitions that most of us have about loving substitutes. Whether this position would be superior to plain a-rationalism is an open question.

7. Kolodny’s relationship account

If we are to solve the problem of substitutes by showing that a transfer of love to the substitute would be not only bizarre, but plain epistemically faulty, we must be prepared either to modify the standard picture or to reject it altogether. There are at least two interesting possibilities to consider. First, one could say that it is not intrinsic properties of the beloved that justifies one’s love but extrinsic ones. Second, one could say that it is (mainly) intrinsic properties as part of a certain context (or, more generally, satisfying certain conditions) that justify one’s love for another. After having introduced and assessed each of these proposals in the next two sections, I’ll outline my own version of the second one.

When considering cases of duplication, one thing is clear: the duplicate is not the person I have been with and loved in the past, she is not the person who had the parents that my partner had, and so on; in short, she is not the person who had the history that my partner had. The duplicate may be qualitatively identical with my partner, having the same body, the same beliefs, the same memories (or ‘quasi-memories’), the same character. As we have seen, this seems insufficient
for me to be justified in loving her instead of the original. What grounds this claim if not the intrinsic properties of the duplicate? A natural thought is the fact that she has a different history.

It is not the mere fact that the duplicate has a different history from my beloved’s that may justify my rejection of the substitute, however. For, it could be asked, why could it not be the case that the fact that the substitute has a different history from my partner’s grounds my coming to love her (the substitute)? There must be something about my partner’s history that makes the relevant difference. So, what has my partner had in the past that the duplicate hasn’t? My partner had certain biological parents, to be sure, but also a childhood, friends, a college education, an art class, a trip to Africa, and a great many other experiences. In all of these, nothing however stands out as the reason for which I love my partner. Perhaps this is because they are all part of the reason. But there seems to be something significant that is missing in the list, namely the relationship that my partner and I have had in the past and that I never had with a duplicate of her.

According to Niko Kolodny (2003), love is grounded, not in intrinsic properties of my beloved, but in his or her relational properties, in particular in the property of sharing a certain kind of history with me. What kind of history is this? Presumably, it is a history of interactions involving certain positive responses (including, as Kolodny calls it, a ‘shared concern’) and the sharing of certain activities. If we are to extend this account to other forms of love, which Kolodny wants to do, biological ties and family bonds may ground love as well. We do justify our partiality (in action and feeling) towards our friends, lovers, and family members by citing the very fact that they are our friends, lovers, and family members. And we surely would find no other answer than ‘Because she is my daughter’ to the question ‘Why do you love this particular girl?’ Does Kolodny give us a plausible account of how love can be justified?

There are clear advantages to the account. First, in broad outline, it can nicely capture not only love between romantic partners, but love between mother and
child, between brother and sister, and between friends. In all these cases, it seems to give an intuitive answer to the question ‘Why do you love him or her?’ Second, naturally, it does not fall prey to the problems of lost properties and substitutes. As we have seen, the mere fact that one has lost some of her valuable properties does not by itself give us reason to consider giving up on loving them. Similarly, the fact that a duplicate of my partner is around should not give me reason to transfer my love to her. Kolodny’s relationship account, by giving non-intrinsic properties the relevant justificatory role, manages to avoid both of these problems. Finally, the account clearly tells us why it is not rationally obligatory to love just anyone who happens to have valuable features. Since it is relationships that justify love, not intrinsic properties, the presence of intrinsic properties in people with whom I have not been in a relationship does not justify my loving them.

Kolodny’s account is too sophisticated to go into the details of it here. However, it is fair to wonder whether the mere fact that one has had a relationship with someone suffices to give one a reason to love or to continue to love the other. What does Kolodny make of the merit intuition? What does he make of the claim that people may not be worth our love? Naturally, Kolodny does not claim that any relationship will do when it comes to justifying our love for someone. Relationships must be of the right sort. For instance, a relationship involving a husband repeatedly abusing his wife does not give the latter compelling reasons to love or to continue loving him. On Kolodny’s account, therefore, positive properties of the beloved—such as her virtues—play a role in enabling a given relationship to become a legitimate source of reason for love. An abused wife, as a result, has reason to stop loving her husband in virtue of the deterioration of the relationship that he causes.

At this point, however, it may be wondered whether, on Kolodny’s view, what really matters to one in a loving relationship is not one’s beloved, but the relationship itself. On the account, the beloved may surely be the intentional object of one’s love—we love people, not relationships—but it is in virtue of the value of one’s relationship with him or her that one’s love is appropriate or
inappropriate. Moreover, given the enabling role that the identity of the lover must play for one’s love to be justified, the lover turns out to be instrumentally valuable. It is as if the content of one’s love could be captured by the thought ‘I appreciate you because you make a worthwhile relationship possible’. If anything, it seems to give the wrong order of explanation: ‘I appreciate the relationship because it enables me to get to know you’ looks like a better thing to say when one expresses one’s love.

In addition, Kolodny’s view seems to fly in the face of our experience of coming to love someone. On his account, one’s love for someone is appropriate just in case it results in a relationship of a certain sort. The object of one’s love thereby becomes valuable to one in virtue of being part of what makes the relationship possible. Yet, it often seems that coming to love someone is more a matter of finding what’s valuable about the other than a matter of finding what’s valuable about the relationship one is engaged in. Of course, we sometimes react to events that bear on the fate of our relationships, as when one of the lovers has to spend some time away for work. But it would be an overstatement to claim that most evaluations manifested by love (on my account, by means of emotional episodes) are about the relationship itself. When we love someone, it seems, we primarily value her as a being that has value independently of the relationship. Our coming to love someone often appears to be more a matter of discovering how wonderful that person is, how worthy of love she is, than a matter of creating an entity that is independently valuable. As a result, claiming that love is an appraisal of a relationship, and of another person only derivatively, doesn’t do justice to much of our experience.

Although we have reasons to reject Kolodny’s specific account, this does not show that history in general, and relationships in particular, have no role to play in an adequate account of justified love. As we will see in the following sections, they do have a significant role to play in the justification of love, and they are crucial in preserving the thought that love not only does not accept substitutes but also is right in doing so.
8. Intrinsic properties contextualised

By now, it should be clear that the standard picture is inadequate as an account of how justified love is possible. For one thing, it seems that the value that love attributes to its object does not depend solely on the object’s intrinsic properties. Moments that I spend with my partner, and activities that we share, are all things to which I am surely not indifferent, suggesting their relevance to my appreciation of her. Furthermore, the satisfaction of common projects—from buying a house to making a painting together—contributes to the individuation of my partner: a perfect intrinsic duplicate of her would not be the person with whom I have produced the relevant things. Kolodny therefore is certainly right to claim that relationships play a crucial role in making one’s love appropriate.

However, we have seen, Kolodny goes too far in thinking that relationships are the primary justifiers of love—i.e., what ultimately justifies instances of love—as the claim leads to the unpalatable conclusion that the object of one’s love is valuable to the extent that it contributes to a valuable relationship. On the contrary, the object of love should be conceived as the primary justifier, and the relationship valuable to the extent that it sheds light on, and perhaps contributes to, the object’s value.

What kind of value does love attribute to its object, then? In some intuitive sense of ‘intrinsic value’, it seems that love attributes intrinsic value that nonetheless may partly supervene on relational properties. On the standard picture, however, intrinsic value supervenes on intrinsic properties alone. Is there any sense to be made of the idea that intrinsic value can partly supervene on relational properties?

Notice that the claim that intrinsic value supervenes on intrinsic properties alone looks more like a substantive philosophical thesis than a simple datum of

101 Relationships indeed typically give rise to the acquisition of new properties in the lovers, both intrinsic and extrinsic.
common-sense. G.E. Moore, the classical proponent of this view, defines intrinsic value in the following way:

To say that a kind of value is “intrinsic” means merely that the question whether a thing possesses it, and in what degree it possesses it, depends solely on the intrinsic nature of the thing in question. (Moore, 1922, 260)

In recent years, several authors have challenged the view that the concept of intrinsic value should be construed under these Moorean lines (Korsgaard, 1983, Kagan, 1998). According to them, an object’s intrinsic value sometimes depends on its extrinsic properties. For instance, Shelly Kagan (1998) suggests that the pen used by Abraham Lincoln to sign the Emancipation Proclamation has not just instrumental value, but by virtue of the role it played in a major historical context, has gained some intrinsic value. Kagan says: “by virtue of that history, we might say, it now possesses intrinsic value: it is something we could reasonably value for its own sake.” (Kagan, 1998, 285)

In order to avoid confusion, I will call ‘final’ the kind of value that opponents of Moore talk about (see Korsgaard, 1983, for the term), leaving aside the question whether it should be conceived as a notion capturing a phenomenon that is distinct from the phenomenon captured by the notion of intrinsic value construed standardly (see Bradley, 2006). I will also leave aside the question whether we are right to think that particulars can be the bearers of intrinsic value, in addition to states of affairs (see Zimmerman, 2001, for a negative answer). What is important to note is that we do seem to treat others as both intrinsically and finally valuable, and that there are standards of justification for such attributions. Treating a morally depraved person who has no valuable qualities whatsoever as intrinsically valuable (for reasons other than that he is another human being) in the way we treat virtuous people as intrinsically valuable looks not merely unwarranted, but outright irrational.
Given that love’s justification may partly depend on the value of the relationship that gave rise to it, it is then plausible to claim that the value that love attributes to its object is *final*. At first glance, it seems that Kolodny could appeal to this notion in order to solve the problems raised above. He could for instance say that love is appropriate just in case the object of love is finally valuable, where being finally valuable in this context supervenes on a historical relation that it has had with me. And given that the object of my love is not treated instrumentally, but for its own sake—a similar way Lincoln’s pen would now be treated—this saves Kolodny’s account from the charge I put forward earlier.

But there is an important difference between one’s beloved and Lincoln’s pen, namely the fact that the former may be independently more or less worthy of being loved whereas the latter has no special value that is independent of the role it played in a particular historical context. In other words, and this is the spirit of the merit intuition, people have qualities that would make them more or less suitable targets of love were a context of the right sort be supplied. The role of Kolodny’s relationships may be to supply this context, and perhaps to provide further epistemic reasons for love (recall the valuable moments spent with our beloveds), but it doesn’t seem to be the case that they *exhaust* such reasons. So, even though love involves the attribution of final value to its object, this final value must in great part (although not entirely) supervene on *intrinsic* properties for love to be appropriate.

To illustrate this point, let’s imagine an individual whose psychology functions in the way a naive situationist (Chapter 5) would think psychology in general functions. All her behaviours, let’s suppose, are determined by both situational factors beyond her control and fleeting mental processes and events that have no deeper psychological source. Imagine now that, after having spent some time with her, you come to love her. Is your love appropriate? According to Kolodny, as long as the relationship between you and her has been going well, the answer is certainly yes. Suppose that most of the situations that you both have found yourselves in include factors that are conducive to positive feeling
and positive behaviour (towards you and towards other people). From your experience of the interactions with your beloved, the relationship is going very well, and you thereby find your love appropriate. Imagine that one day you realise the truth about your beloved’s psychology. You realise for instance that, even though the interactions have been positive, they in part result in your beloved’s having reacting stereotypically in ways that you find congenial. Would you now consider yourself as justified in loving your partner? Perhaps you would—after all, we may have a natural inclination to think of ourselves as minimally justified in loving the people we love. But would you consider yourself as as much justified in loving her as you used to? My suspicion is that the discovery of your partner’s actual psychology would at the very least make you doubt that she is as worthy of your love as you thought. If this right, then relationships, even good ones, are not sufficient to fully justify love for particular (previously unrelated) people.

The fact that I have been in a relationship with my partner, and not with her duplicate, makes a difference to how I am justified in loving one or the other. But if this difference is not primarily to be explained in terms of the relationship playing the main justificatory role, how is it to be explained? As hinted at earlier, relationships furnish the context in which intrinsic properties become epistemically significant. By involving many interactions and shared activities, they enable the participants to have access to facts about each other that may not be easily available to outside parties. What are these facts? Presumably these facts include facts about physical appearance, such as the presence of a beauty spot on one’s beloved’s toe. But as clear from the previous thought experiment, these facts should also include deeper facts about the other’s psychology; not merely her thoughts, emotions, appetites, and so on, but what drives these mental states, namely her personality, her cares, her character.

102 And in which the valuable relational properties described above are made possible. I will however put them aside in the rest of the discussion, as it should be clear from the previous thought experiment that the value of these properties is largely dependent on the value of the intrinsic properties of the participants of the relationship.
Intrinsic psychological dispositions, therefore, appear to be the right sort of things to count as justifiers. And intimate relationships are plausibly one of the contexts where the manifestation of psychological dispositions is most likely. By spending time with someone who loves us and who has made themselves vulnerable to us (by putting aside various inhibitions), we put ourselves in a privileged epistemic position in which we can witness deep facts about the other’s psychology reveal themselves. By seeing how they react to various events, we have an indication of what they care about and what their character is like. We can infer how compassionate, kind, and generous they are (although we tend to pay less attention, we can also infer how imperfect they are). Of course, such ‘general’ or global character traits may not be sufficient for love to be justified. An abusing husband who is kind, compassionate and generous towards everyone except his wife surely is not very worthy of her love. Are general character traits necessary for justification? This is a matter for dispute, but there seems to be something appealing about the claim that for X to be worthy of Y’s love, X must manifest at least local traits that are good for Y such as kindness-towards-Y, generousness-towards-Y, desiring-to-be-close-to-Y, being-concerned-about-Y’s-wellbeing, and so on. The manifestation of these dispositions in X is plausibly necessary for justified love in Y. I will remain silent on whether it is sufficient, although I suspect that, although it may be sufficient for a love that is more appropriate than not, it may not be sufficient for a love that is highly appropriate. (Consider a Ted Bundy-like individual who, although very morally depraved in most respects, is a loving and caring husband.) As long as it is clear that more general cares and character traits typically do play a significant role in the justification of one’s love, we need not be concerned about this issue.

To recap, love can be appropriate in virtue of various properties of its object. These properties include relational/historical properties, but are not exhausted by them. Certain intrinsic properties of the beloved must ground love as well, and even seem to be the primary justifiers of one’s love. One’s relationship with the other, when not providing independent epistemic reasons for one’s love,

103 What Abramson and Leite call ‘virtues of intimacy’ (Abramson & Leite, 2011).
works as an enabling condition for the relevant intrinsic properties to provide one with reasons to love, or to continue to love, the other. The difference between loving my partner and loving an intrinsic duplicate of her thus lies in the fact that my partner’s properties have manifested themselves in a relational context that enables her properties to count as reasons for my love in a way that the duplicate’s properties don’t.

One might wonder whether the problem of substitutes has really been solved with such an appeal to contextual parameters that supposedly make an epistemic difference. For one thing, we have accepted the claim, which I dubbed the merit intuition, that an individual who possesses certain intrinsic properties may be worthy of being loved. So if my partner is worthy of being loved in virtue of her intrinsic properties, an intrinsic duplicate of her is likewise worthy of being loved (and I know it). So why is it not the case that I have good reasons to love the duplicate? What is it about the fact that I have been in a relationship with my partner that would make loving an intrinsic duplicate of her epistemically faulty? What’s so special about the relational context? In the following section, I will provide the outline of an answer, giving us the last elements for an adequate rationalist account of justified love.

9. Synchronic and diachronic justification

In order to know what is special about relationships that would make loving a substitute inappropriate, I suggest, we must be clear on the nature of the justification that one typically has when one has been in a loving relationship for a significant period of time and that one fails to have when one hasn’t been in such a relationship. The claim is not that, without the relevant relationship in place, one has no epistemic reason to love anyone else—as we have seen, this is something we are not prepared to accept. The claim is rather that the kind of justification that one has when part of the relevant relationship is different in some respect from the kind of justification that one has when not part of that relationship.
More crucially, the kind of justification that one has when part of a newly formed relationship may be different in some important respect from the kind of justification that one has when part of a more mature relationship. Given the nature of the properties that typically justify one’s love, it is unsurprising that it takes time for one’s love to reach the state of appropriateness found in certain long-term relationships. In other words, it takes time for one to be justified in loving someone else. And this is surely something that the antirationalist gets right when she claims that most instances of love are not formed on the basis of compelling reasons, and that the irrationalist may be right to emphasise when she claims that a great many instances of love are outright inappropriate (given that, according to the rationalist, many of them do not reach the relevant state of maturity).

What the antirationalist gets wrong, however, is the possibility that one’s love become justified over time. Given the nature of love’s justifying properties, this possibility is not incredible indeed. Let’s call synchronically justified any attitude that is justified at a given time $t$ and diachronically justified any attitude that is (i) synchronically justified at $t$ and (ii) owes its being synchronically justified at $t$ to an adequate process of inquiry or experience (where by ‘adequate process’ I mean ‘process without which the synchronic justification of the attitude would be significantly lower’). Thus, while an attitude can be synchronically justified at the moment of its inception, it is diachronically justifiable just in case further inquiry or experience can in principle make it more (or better) synchronically justified.

A tentative example of an attitude that is synchronically justifiable but not diachronically justifiable is that of perceptual belief. When I see a red object, in good lighting conditions and with no prior reason to think that my visual capacity is in any way impaired, I can form the belief that there is something red in front of me (or if even that state’s justification can be improved over time, I can form the belief that there is something that seems red to me). In such a case, there seems to be no way the belief could become more justified than it is at the moment it is formed. Examples of diachronically justifiable states are easier to
find. One’s belief in a scientific hypothesis for which some but not overwhelming evidence is available but that could in principle be supported by further evidence may be one such example. Belief in God and various moral beliefs may be good examples as well. The distinction between synchronic justification and diachronic justification is therefore far from a convenient ad hoc construct.\textsuperscript{104}

If love is an attitude that is diachronically justifiable, we should expect it to be the kind of thing that can be formed on the basis of less than overwhelming evidence. (This may explain why it may be rationally permissible as opposed to rationally obligatory to love some particular people.) We should also expect it to be something for which we can be more or less justified at different moments in time, depending on the details of its history. Finally, we should expect that the longer the history one’s love has, the more likely it is to be epistemically justified (or, perhaps, unjustified). I think these expectations are right on target. First, as the irrationalist likes to point out, love is often based on little, bad and perhaps even no evidence. Note however that most of the cases she cites are cases of newly formed loving relationships. Second, given the history required to have access to the relevant justifying properties, it is no wonder that love can not only vary in justification but be \textit{more} justified over time. (Of course, it can also decrease in justification if the expected properties turn out to be non-existent or if the beloved has lost them in the meantime.)

How is the distinction between synchronic and diachronic justification going to help us solve the problem of substitutes? Although it will not do all the work, and will have to be supplemented with a further assumption, I think it can give us an account of the difference between justifiably loving someone as a result of a process of inquiry or experience and justifiably loving someone as a result of an abstract awareness of the properties in virtue of which one is worthy of love.

To do so, let’s first distinguish between \textit{new loves}, which are instances of (let’s assume, minimally synchronically justifiable) love formed in newly formed

\textsuperscript{104}See Swinburne (1981) for the distinction in the context of religious belief.
relationships, and *mature loves*, instances of diachronically justifiable love. Let’s say X *loves* Y just in case X has a new love for Y, and X is *justified* in loving Y just in case X possesses evidence that makes her love for Y minimally synchronically justified (where minimal synchronic justification is justification based on small or at any rate less than overwhelming evidence). Let’s further say that X *loves* Y just in case X has a mature love for Y, and X is *justified* in loving Y just in case her love is diachronically justified.

Now, we can see why it seems true that, although people that we don’t love can be worthy of love, we are still in some sense not justified in loving them. What we mean by this is that, although other people may be worthy of being loved, we have no reason to love them because to justifiably love them requires first going through the required process of inquiry or experience constitutive of justification. At best, therefore, we are justified in loving, but not loving, other people. However, given that appropriate instances of love are based on justification, that is, less than overwhelming evidence, we are only rationally permitted to love them, which seems to be the right result.

In cases of duplication (at least where the duplicated partner is one I love and am justified in loving), it may seem that I am justified in loving the duplicate, for she has the same properties that my partner has and that provides me with the justification to love her. The problem, however, is that this seems to be wrong. So there must be a story that explains why while I may be justified (that is, justified) in loving the duplicate, as I may be justified in loving virtually anyone, I am not justified in loving her. Notice that the distinction between synchronic and diachronic justification (and between new love and mature love) allows us to accept that the duplicate may be both worthy of love and love, although it would be inappropriate *for me* to love her, for to love I need to go through the relevant process. But if the process’s role is to give me access to my partner’s justifying properties, then why can’t the fact that it also gives me indirect access to the justifying properties of her *duplicate* be enough for a love for the latter to be fully appropriate? Again, what’s special about our relationship? At this point, it may be wise to say that rationalism cannot solve
the problem of substitutes, at least when the relevant substitute is stipulated to be an intrinsic duplicate, and that our reluctance to love the duplicate may be due to various a-rational factors, or, à la Kolodny, relational properties of my beloved. And I agree that it may be the way to go for the rationalist if no adequate response to the problem is available. However, I think an adequate response is possible.

Suppose that you were to meet an intrinsic duplicate of your partner. The duplicate, as a result, would have the exact same appearance and psychology as those of your partner. Or so those having made the duplicate are telling you. In this situation, you may well believe that the duplicate possesses all the dispositions that you value in your partner, but the evidence that you have in favour of this belief includes her appearance, her demeanour and behaviour, and other subtle cues that may crop up in your encounter with her. It is not however the sort of evidence that you have had the occasion to gather over the course of a few years of relationship with your partner. Moreover, the fact that the duplicate’s makers tell you that all the properties that you value in your partner are also properties that the duplicate possesses may constitute good grounds for thinking that the duplicate is very worthy of love (or ‘love_2’), but it is nothing compared to the years of active engagement that you have spent with the original. The kind of evidence that you have for loving the duplicate, as a result, only gives you reason to love_1 her, and that even though we may be justified (i.e., justified_1) in believing that she is perfectly worthy of being loved. Why it would be faulty to love_2 a duplicate of my partner is simply the same reason as why it would be faulty to love_2 just about anyone whom I believe to be worthy of love_2: that I have not gone through the relevant process with her. 105 What’s special with relationships, therefore, has to do, not just with the fact that it gives one the opportunity to have access to justifying properties, but with the particular, chiefly first-personal way they are accessed.

105 Except of course that I am much more confident that, were I to have a relationship with my partner’s duplicate, I would be more likely to be justified_2 in loving her.
Of course, it may be wondered what is so special about this first-personal way of accessing our beloved’s valuable properties that would make loving (that is loving2) any substitute inappropriate. I admit that I do not have an answer to this question. I suspect however that this is a particular case of the more general problem of why there seems to be something wrong with the idea of evaluative attitudes being justified by non-experiential means such as testimony. To give an example, although we may be justified in thinking that a given film is probably excellent on account of a friend’s suggestion, there seems to be something fishy about actually judging the film as excellent on that basis. Similarly, when claiming that eating meat is wrong, we typically don’t want others to agree with us blindly (as we may agree with what scientists say), even though we may think of ourselves as perfectly justified, but expect them to think about the issue for themselves and arrive at a conclusion on the basis of this thinking. What precisely is missing in these cases is a matter of dispute (see, for instance, Hopkins, 2007). However, by showing that our solution can rely on such broader considerations, we thereby show that it is not entirely ad hoc.

None of this is to suggest that our strong reluctance to love a substitute can wholly be explained in epistemic terms. We have indeed accepted the possibility that we have a non-rational inclination (or what I called an attachment) to continue loving and caring about the people we love and care about and not to accept substitutes. What the rationalist has shown is rather that, whatever other reasons we may have not to love substitutes, we can also have epistemic reasons not to do so. Loving particular people, as a result, can be rationally grounded.

In this section, I have argued that love for particulars should be conceived as an attitude that can be justified not only at a time, but also over time. So, although ‘new’ love may be based on very little grounds at its inception, nothing prevents it from acquiring strong justification over the course of a long-term relationship. Starting to love someone is very much like putting a bet on that person; it is being confident that she has what it takes to fully justify one’s love. Of course, disappointment is often around the corner, giving the impression that love is always blind, mistaken, or irrational. The rationalist account I have put forward
readily accounts for, and in fact predicts, this fact of life. At the same time, the account provides a story as to how, besides all the instances of irrational love, justified love is possible. Rationalism about \textit{de re} love between previously unrelated people is therefore vindicated.\textsuperscript{106}

\textbf{10. Conclusion: The possibility of refined emotions}

In this final chapter, I have argued that, although not all forms of caring and love are subject to epistemic assessment, many are. Not only can it be appropriate to care about certain kinds of things and not others, it can also be more appropriate to care about some tokens and not others of the same kind of thing. One way we can care about some tokens of a type and not others is impersonally, by caring about the relevant tokens, whatever they are, because they satisfy a certain description or play a given role. If rationalism were true only of such ‘\textit{de dicto}’ caring, this would however be a cheap victory, as this form of caring is far from paradigmatic and may not even look like genuine caring at all. Caring for particulars, of which love is the paradigm, is what the rationalist must account for in order to win the battle. Fortunately for her, we have shown that rationalism can account for this particular form of caring (at least between previously unrelated persons).

Here ends our defence of sentiments. We saw that sentiments can be brute causal forces that dispose us towards various emotions. We also saw that they can be attitudes, and that cares constitute an important subset of them. Cares, moreover, can come in various forms. They can be desire-like in representing the world as to be a certain way, or they can be belief-like in purporting to represent the world as it really is. Among cognitive forms of caring, we have those that are utterly irrational, as in blind love, and that it may sometimes be a good idea to get rid of; but we also have those that can acquire a proper rational

\textsuperscript{106} Can the account be generalised to Frankfurt’s favourite case, namely love for our children? I suppose the rationalist could attempt to show that her account can be generalised. However, I am uneasy about this possibility, as it would seem to imply that a mother’s love for her newly born infant is not fully justified, which I take to be a \textit{reductio} to this generalising move.
grounding over time, and be as a result things that we can cultivate and cherish—‘mature’ love being one such case.

As we have seen, given that sentiments are distinct from the emotional episodes that they produce, the possibility that sentiments acquire epistemic maturity (or ‘justification\textsuperscript{2}’) over time opens up another interesting possibility: that the emotional episodes can owe part of their justification to the sentiments that produce them. Deonna and Teroni (2012, 114) are sceptical. They claim that, given that sentiments (of the cognitive sort) owe their justification to the series of emotional episodes that produced them, they cannot in turn contribute to the justification of the emotional episodes that constitute their manifestations. And if it were the case that sentiments cannot improve in justification over time, then Deonna and Teroni might be right in claiming that sentiments play no justificatory role with respect to the emotional episodes that they produce. However, as I have argued, sentiments can be justified in the relevant way. As a result, there is theoretical room for pursuing the hypothesis according to which emotional episodes can become more and more justified in virtue of being produced by more and more mature sentiments.\textsuperscript{107} Perhaps, indeed, this is what it takes to have a ‘sentimental education’.

\textsuperscript{107}See appendix for more details.
Appendix

A Reply to Deonna and Teroni (2012)

In their recent introduction to the philosophy of emotion, Julien Deonna and Fabrice Teroni express some doubts about the idea that what they call ‘motivational states’, such as character traits, sentiments, and desires, can contribute to the justification of the various emotional responses we undergo (114-115). They consider two conceptions of these states, one taking them to be distinct from emotions but nonetheless connected to them in some way (roughly the view I defend in this thesis) and one taking them to be constituted by emotions (e.g., Helm, 2001). They then argue that none of these conceptions can help us make a plausible case for the possibility of a positive epistemological contribution—that is to say, a justificatory contribution—of motivational states to the emotional episodes to which they are related (either by the disposition-manifestation relation or by the part-whole relation).

Against the realist about motivational states (i.e., the one who claims that such things are distinct from emotional episodes), they argue that, even if we take such states to be “long-standing intentional relations” (what I call ‘attitudes’), and therefore if we allow them to be justified, it is unclear that their justification can play a genuine role in the justification of the emotional episodes that they produce. Two reasons are given to support this claim. First, motivational states would seem to owe much of their justification to the pattern of emotional responses that gives rise to them; the justification of these states, in other words, depends on the justification of prior emotional episodes. “Intuitively”, they say, “it makes more sense to say that my love for all things Shakespearian inherits its justification from the many specific and distinct emotional experiences his works have elicited in me, and out of which my love has emerged, than the other way around.” (114) The second reason is the following: “the advocate of this view will have to explain away the intuition that many emotions—like the
admiration one may feel on discovering an artist’s work—do not seem to be accounted for by any long-lasting intentional relation (...) and a fortiori by one from which they can inherit their own justification.” (ibid.)

On the second conception of motivational states Deonna and Teroni mention, emotional episodes depend on motivational states for their justification because such motivational states, of which the emotional episodes are constituents (à la Helm), partly or wholly constitute the values (or formal objects) to which these episodes respond. “Because the relevant motivational states partly constitute the evaluative properties to which emotions respond, one is justified in responding with a given emotion only insofar as one is in the relevant motivational state.” (115) Again, they provide two reasons to reject this possibility. First, they point to another place in the book where they provide what they take to be principled reasons to think that all forms subjectivism about values (of which the view under consideration is one) are mistaken. Second, they observe that the fact that we commonly appeal to motivational states in order to answer questions such as “Why does she feel this way?” does not support the sort of subjectivism at issue. “These answers”, they say, “appear to have the function of explaining why the relevant emotion occurs rather than that of explaining why it is justified or correct. This is especially clear in all those cases we have discussed in which motivational states are conscripted into explanations of why unjustified emotions occur.” (ibid.) (I take it that this last argument can also be formulated against a non-reductionist or realist about motivational states who also takes such states to constitute value.)

In the following, I will briefly introduce plausible ways these arguments—which I suspect Deonna and Teroni take to constitute challenges rather than the last word on the matter—can be met if we accept my account of sentiments. First, I will suggest ways a realist about motivational states who takes values to be independent from them (in accord with Deonna and Teroni’s own metaethical view) can nonetheless allow such states to play some role in the justification of the emotional episodes they produce (as hinted at the end of Chapter 6). Second, I will suggest that a realist about motivational states has the
resources to defend the claim that at least some values can depend on such states, thereby opening up the possibility that such states contribute to the justification of emotions, in a way that bypasses the problems Deonna and Teroni raise against subjectivism (although it would remain to be seen that the account avoids all problems that various forms of subjectivism face). It should be noted that nothing I will say is supposed to establish any particular view about the justification of emotions. Nor are the considerations I will give supposed to be taken as fully worked-out. My aim is only to show that there is no reason to suppose, prior to inquiry, that sentiments realistically construed, and as (partly) cognitive attitudes that can be justified over time, cannot contribute to the justification of the emotional episodes they produce.

Realism about sentiments + sentiment-independence account of values

This first combination of views is the most congenial to Deonna and Teroni’s own views, as they both accept realism about value and may be happy to accept my account of sentiments (at least in broad outline). Since they take emotions to be responses to value properties instantiated in the world (i.e., the formal objects of emotions), however, they are reluctant to make the justificatory status of emotions depend on mental states other than those that supply them with the necessary descriptive information (chiefly perceptual and belief states—what Deonna and Teroni call the emotions’ ‘cognitive bases’), that is, information about—on their account—the supervenience base of the relevant value properties (e.g., the dog’s big teeth for dangerousness). On such a view, it is difficult to see what justificatory role, if any, sentiments could play in addition to the role played by the emotion’s cognitive bases. As Deonna and Teroni say, sentiments may at best be justified by the emotional episodes that created them, which casts doubt on the claim that sentiments can in turn justify the emotional episodes that they produce.

One problem with this last argument is that from the claim that mental state S owes its initial justification to pattern of mental events P does not follow the claim that S cannot in turn justify further mental events (in particular the ones
that it produces) in the future. For one thing, the presence of a stable disposition towards the relevant emotional episodes is evidence for the fact that the attitudes in question are not correct or fitting only by accident. If one accepts, as I do (Chapter 1), the metaphysical possibility of an entire pattern of emotional episodes that all turn out to be correct (i.e., the formal objects of the relevant episodes really obtain in the world) but that are induced by an external factor, one by one, then one must ask whether the relevant emotions can be said to be as justified as they would be if produced by underlying (justified) sentiments. If the answer is no, then we have a reason to think that sentiments can in principle contribute to the justification of the emotional episodes that they produce. Note that this is quite in line with the distinction I drew in Chapter 6 between (merely) synchronically justified sentiments and diachronically justified ones. We saw, for instance, that the difference between a ‘new’ love and a ‘mature’ love lies in the fact that the latter, and not the former, is supposed to have stood the test of time, so to speak, by having acquired strong justification for the evaluation it makes. The details about the nature of this justification remain to be filled out (perhaps, for instance, the fact that the sentiment coheres with other mental states may constitute part of the story), but as long as it is clear that a pattern of emotional episodes triggered by a (justified) ‘new’ love may overall be less justified than a pattern that is exactly similar but that is triggered by a (justified) ‘mature’ love, then we are on our way to establishing the view that Deonna and Teroni reject. I won’t pursue this line of thought any further, but I suspect that a plausible character- (and therefore, on my view, caring-) oriented account of justified emotion can be constructed.

It may seem however that the kind of account of justified emotion that I am suggesting does not accommodate Deonna and Teroni’s case of single emotional episodes that seem to be justified independently of whatever motivational states the subject happens to have. For instance, the admiration felt as a result of discovering an artist’s painting (when, let’s suppose, one had no prior interest in the arts) seems justifiable by appeal to the ‘admirability-

---

108 Sentiments may for instance be required for reliable knowledge about value (suggesting some sort of ‘tracking’ account), or for responsibility in the formation of justified emotions (ruling out alienating emotions as unjustified or at any rate less justified), or both.
making’ features of the painting. These sorts of cases appear genuine, and their possibility may even be required for genuine changes in evaluative outlook to occur. Such reactions indeed sometimes feel like pure revelation of something we previously not only didn’t care about but also were completely unaware of. What are we to make of these possibilities? Notice first that the kind of account of justified emotion that I suggested above does not deny that emotional episodes can be justified to some extent irrespective of their dispositional origin; it only claims that the latter can contribute to their justification. But even if it claimed that the relevant emotion is not justified, the account would have the resources to show why, namely that it is purely by accident that the emotion got the value property right. Being only accidentally fitting, the emotion may as a result lack the sort of justification that, say, the admiration of someone with a developed artistic sensibility would have towards the same painting.

An alternative way to explain the isolated emotion case is to appeal to broader concerns that the subject has. We can grant that, in the case under consideration, the subject’s emotion is not based on any sort of developed artistic sensibility or concern for good artworks. This however does not force us to think that the emotion as a result has no deeper source, or that it is directly imposed on us by the beauty of the painting (via the cognitive bases), or some such. Of course, this could sometimes be the case, in which case the above story holds. But this need not. Broader concerns, such as one’s caring about beautiful things in general, may certainly explain the admiration one feels upon seeing an artwork for the first time, and, insofar as they can themselves be justified in the way suggested above, they can in principle contribute to the justification of isolated emotions of this sort.

As far as I can tell, therefore, nothing Deonna and Teroni say casts doubt on the claim that, assuming a sentiment-independence account of values, sentiments realistically construed can in principle contribute to the justification of the emotional episodes that constitute their manifestations.
I think that Deonna and Teroni’s response to the second account, although not entirely unsatisfying, is not as strong as they make it out. The two reasons they give, we have seen, are (1) that subjectivism about value is implausible and (2) that the common appeal to sentiments to explain emotional episodes cannot be used as a reason to favour subjectivism about values (or the formal objects of these episodes). In the following, I’ll grant (2), as I think Deonna and Teroni are right to claim that our emotions are typically not justified by an appeal to their dispositional origin—if anything, we justify them by appeal to (intrinsic or extrinsic) properties of their object. This however does not mean that the dispositional origin of our emotions cannot constitute the relevant values in any way. They could for instance provide the background condition required for certain objects (properties, facts, events, states of affairs, etc.) of the world to count as valuable or disvaluable. They could also be part of the supervenience base of the formal objects of our emotions, namely part of relational properties of their objects. For instance, if the formal object of sadness is \textit{loss of something dear to one}, then it is unsurprising that the thing really must have been dear to one in order for the emotion to be fully justified. Whether or not, on such accounts, sentiments can be said to literally justify our emotions, it is clear that they at least make fully justified emotions possible.

Now, the reasons Deonna and Teroni give to reject subjectivism about value are rather well known and do not need to be developed in full detail here. After discarding implausible radical forms of subjectivism (e.g., the thesis that to be valuable is simply to elicit certain emotional responses), namely views analysing value properties in terms of actual responses, Deonna and Teroni consider views that take values to be constituted by \textit{dispositions}. One such view

---

109 Perhaps the background condition would be required for certain objects to have \textit{intrinsic} value, i.e. value that, as long as the condition holds, supervenes on intrinsic properties of its bearer. For this possibility, see Rønnow-Rasmussen, 2011, Ch.2.

110 One reason these views are implausible is that they have the unpalatable consequence that we cannot ever be wrong in attributing value properties to objects, as long as we have the relevant emotions towards them.
has it that an object instantiates a given value (e.g., admirableness) just in case it is disposed to produce (say) admiration in us. This view has the advantage of allowing cases of incorrect emotions as, on such a view, “the exemplification of evaluative properties is independent of any particular occurrence of an emotional response.” (43) At this stage, Deonna and Teroni suggest, the account remains very sketchy. As they say:

A fully fledged version of dispositionalism must reach a decision as to which subjects should be appealed to, and, among their responses, which are to count as manifestations of the relevant dispositions. Does the admirable character of a work of art depend on its disposition to cause admiration in all and every creature, or only in those that show a minimal interest in the fine arts, or perhaps only in established experts? (43-44)

Of course, the subjectivist who takes sentiments to be distinct from the relevant emotional responses may say that the relevant class of people here is that of those who (say) have a genuine concern for artworks, or for objects that tend to produce positive emotions (such as admiration), or some such. And she may even allow certain sentiments, especially those whose existence depends on the existence of other mental states (including deeper concerns), to be justified (in line with much of the argument of Chapter 6), although the details of this view remain to be filled out (see Goldman, 2009 for a view roughly along these lines). In any case, there seems to be nothing arbitrary in the decision to select the class of people who care as the relevant class of subjects where values (i.e., dispositions of objects) can manifest themselves.\footnote{Given that I take dispositions to be genuine properties of objects, I can’t help seeing this view as a species of objectivism about values where values can be accessed only via their manifestations (i.e., our emotions) and, if they are to be reliably accessed, only by certain people (i.e., those who care about these dispositions). But from what Deonna and Teroni say in the following, it is clear that they construe dispositionalism as a view where our own dispositions partly constitute values.}

111
But there is a deeper worry about subjectivism that Deonna and Teroni have in mind, and which, according to them, should lead us to reject it in all of its forms. They write:

Note (…) that a very problematic upshot of this approach is that the world could become a better place (with fewer evils) if future generations were simply inured to torture, poverty, slavery, etc., which are per hypothesis evils only due to our current indignant sensitivity to them. (44)

They conclude that we should look for an account of value that agrees with “the idea that the existence of a widespread and stable disposition to have a certain emotional response is compatible with the absence of the relevant evaluative property.” (ibid.)

Notice that Deonna and Teroni’s target is now not the initial dispositional theory according to which values are dispositions of objects to produce emotional responses, but one according to which dispositions of persons constitute (at least in part) values. If they were still discussing the former account, they could still admit that, in a certain sense, values could still exist, albeit in unmanifested form, even though we stopped being sensitive to them. The view that they are criticising is rather that, without the relevant sensitivities, values would not exist at all, which surely looks genuinely subjectivist.

I’d like to accept the thought that making the (dis)value of things such as poverty, slavery, etc. depend on, as Deonna and Teroni call them, ‘brute psychological facts’, doesn’t seem right. I’ll therefore accept the objectivist’s contention that, if such values (disvalues) exist, they do not depend on our sensitivities. I prefer to say ‘if’ because I would like to remain silent on the question whether moral value actually exists, and therefore whether an error theory about it is right. But I take it that one can be an objectivist about moral value—like I would be if I were convinced that it in fact exists—and be a subjectivist about other kinds of value. Consider for instance a drawing that my
little cousin drew a few years back. Let’s assume that it is not clearly better than
any other drawing from someone of her age. We can further hypothesise that the
paper has worn out and that the drawing is barely visible. Arguably, the sort of
value, if any, that it has would not be the sort of value the objectivist has in
mind when she talks about moral value. In addition, it doesn’t seem to be of the
sort of value the objectivist may attribute to a great artist’s paintings. Still, the
drawing has value for me; to use a philosophical term of art, it has personal
value. The question, then, is whether we should be objectivists or subjectivists
about this kind of value, irrespective of our view about the nature of moral (and
more generally impersonal) value.

Deonna and Teroni, therefore, cannot use the arguments they give against
subjectivism in order to build a case against the view that our sentiments (in
particular our cares and concerns) constitute personal values. And this is
precisely the kind of view that their main target, Bennett Helm, defends in his
values, are values that are relative to the individual in the sense that what
personal values it is right for me to hold may well differ from those it is right for
you to hold” (12). The view he defends at great length is based on the thought
that to care about something is to instantiate a pattern of rationally
interconnected emotions that have a common focus (see Chapter 1). Each
individual emotion, moreover, is justified only if it is part of such a pattern.
Personal values, he suggests, should be conceived as constituted by the subject’s
cares (hence by the relevant patterns).

It is not the place to assess Helm’s view (although we have seen that caring
should not be reduced to patterns). The point is rather that Deonna and Teroni
have not yet given us any reason to reject a subjectivist view about a certain
class of values, and therefore the claim that some of our emotions are justified
in virtue of being part of our cares (or, in my view, in virtue of originating in
our cares).
Of course, it remains to be seen whether this account of personal value is adequate, and perhaps how it can be combined with an objectivist account of impersonal value (avoiding therefore the objections Deonna and Teroni raise). But, as far as I can tell, there is no reason to reject it prior to systematic inquiry.

\footnote{I suspect this would be one way to square subjectivism about value with the considerations given in Chapter 6 (in particular what I called the 'merit intuition'). Although his paper is about normative reasons, and not values \textit{per se}, David Sobel would probably call this combination of views 'weak objectivism' (Sobel, 2005, 437).}
Bibliography


Bates, T. (manuscript). “In Defence of Vice”


Choi, S. (manuscript). “Purely Dispositional Worlds”


