LEARNING ABOUT ATTACHMENTS THROUGH LITERATURE

Radu Bumbăcea

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

The University of Leeds
School of Philosophy, Religion and History of Science

APRIL 2023

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

ACKNOWLEDGEMENTS

This thesis is heavily indebted to my supervisors, Víctor Durà-Vilà, Matthew Kieran and Helen Steward. I started with Matthew Kieran as a primary supervisor, and I have benefited enormously from his no-nonsense approach to philosophy and his sharp questions. Matthew retired at the end of my second year of PhD, but luckily what could have been a difficult situation has actually become an opportunity to work with Víctor Durà-Vilà, who stepped in to become my new primary supervisor. I have very much enjoyed working with him and have learned a lot about how my work relates to or contrasts with other ideas in aesthetics. Víctor has also insisted that I rewrite and restructure bits of my thesis, and now I can see that he was right. Saying that Helen Steward was my ‘secondary supervisor’ does not do justice to her contribution to the thesis, especially towards the first half, which is more about philosophy of mind. Through successive meetings, she has greatly helped me develop and sharpen some key ideas of the thesis, which, even though I felt they were on the right track, were still muddled.

In the third year of the PhD, I spent three very pleasant months in Geneva, where some of my work has been read and insightfully commented on by Julien Deonna, Julia Langkau and Fabrice Teroni. Those months have helped me develop my views on emotions and I suppose their result is readily visible in the thesis. The placement in Geneva has been partly supported by a Turing Scheme grant, for which I am grateful.

The PhD culminated with a very instructive viva with Julian Dodd and Derek Matravers as examiners, whom I’d like to thank for their thorough engagement with my work. The viva helped me see many avenues for the future that would continue the work done in the thesis.

Some of the ideas from the first and last chapter pre-date my PhD, having been incipiently developed in previous work in Cambridge and London with Nick Denyer, Stacie Friend and Andrew Huddleston. The first chapter served as my transfer chapter and has later been improved in light of very useful suggestions made by my transfer examiners, John Divers and Jessica Isserow.

Many friends and colleagues have helped me with comments, advice and encouragement: Agnès Baehni, Elodie Boissard, Constant Bonard, Aleksander Domosławski, Jakob Donskov, Simon Graf, Steve Humbert-Droz, Roberto Keller, Sam Mason, Antoine Rebourg and Leonard Weiss. They have been a wonderful part of my PhD life.

Lastly, I'd like to thank my parents, Roxana and Dragoș, who have supported me throughout these years.

ABSTRACT

This thesis defends a theory of literary cognitivism: some good literary works, novels in particular, acquaint us with remarkable attachments and put us in a position to evaluate them, and thereby increase our ethical knowledge significantly.

Chapter 1 argues that an attachment, such as those involved in love and friendship, is a drive, that is, a standing mental state that manifests itself in emotions and encapsulates what one ultimately cares about. This drive is directed at the person (or entity) one is attached to. Chapter 2 defends a version of the attitudinalist theory of emotions of Deonna and Teroni (2012), that emotions are bodily attitudes directed at a content. It also argues that emotions which are manifestations of attachments do not have fittingness conditions. Chapter 3 provides a theory of imagining emotions, that imagining an emotion is forming a thick meta-representation of that emotion, where a representation is thick if its object can be apprehended in the representation.

Chapter 4 argues that literature can provide knowledge what an emotion is like by helping us imagine it. However, this does not amount to understanding the attachment that such an emotion might be a manifestation of. Chapter 5 thus defends the main claim of the thesis, that a work of literature can help us imagine relevant emotions of a character and put them together, in such a way as to acquaint us with an attachment of that character and allow us to judge it ethically. Chapter 6 accounts for why this is an important epistemic gain. The following principle is defended: the value of an individual attachment is revealed in that instance and is only minimally connected to a general description. General judgments of attachments are based on assemblages of individual judgments, so can be changed by an acquaintance with remarkable instances.

TABLE OF CONTENTS

INTRODUCTION

As readers, we often say that we are inspired by Fanny Price's devotion to Edmund in Austen's *Mansfield Park*, Maggie Verver's care for her father in James' *The Golden Bowl*, Alyosha's love for his less than perfect family in Dostoevsky's *Brothers Karamazov*, or Newland Archer and Ellen Olenska's commitment to the New York society they belonged to, a commitment that leads them to the greatest sacrifice, in Wharton's *The Age of Innocence*. What these examples have in common is that they involve what might be called an 'attachment', whether to a person, a family or a society. Being inspired by these attachments, we might be tempted to say that they have changed our life in some way. On a first hearing, all this might sound like being inspired by one's older sister who made it into the Bar and hence showed that 'it can be done' and so that 'I can do it as well'. In such a case, the inspiration consists in giving one some impetus to realise a plan that was already in one's mind, an impetus derived from seeing the plan realised by someone else. But this is far from what happens when readers are inspired by some literary character's attachments. What seems to happen, and what I will argue in this thesis does actually happen, is that in reading one of these works of literature, we understand the attachments that the characters instantiate and come to judge them – the work acquaints us with the attachments of the characters and puts us in a position to evaluate them. Such an acquaintance with a remarkable attachment and our evaluation of it, I will then claim, amount to a form of *learning*, of increasing our ethical knowledge.

This is a thesis about what we can learn from literature, about the role that literature can play in our life. It is safe to say that it is a defence of a fairly strong version of the position that literature is highly important in ethical understanding. Before going into the details of the thesis, I'd like to sketch what lies behind it, what motivated some of the thoughts in it. There are two driving interests of mine that have shaped the thesis.

The first one is in ethics, and in particular in attachments, mostly to other people, but also to groups, places, etc. Following some influential papers (Stocker 1976; Williams 1981; Wolf 1982) that have complained that, amongst other things, ethical theories cannot account for the importance of attachments, there has been a surge in debates about love and friendship. Even if many good points have been made about the nature of attachments (Rorty 1986; Cocking and Kennett 1998; Jollimore 2011), I was not really content with any view on the market of what exactly attachments *are*, so I will put forward an alternative one.

Although I will build a theory of attachments and of how they relate to our emotions, a theory that will involve some heavy-duty philosophy of mind, this theory will be more like a framework, by which I mean that it is the starting point, rather than the end point of understanding and evaluating attachments. Furthermore, as will transpire by the end of the thesis, I believe there's only so much that we can theorise in this field, a view that might place me, in spite of major methodological differences, in a line of philosophers that have complained about moral philosophy's tendency to over-theorise (Nietzsche 2006 [1887]; Williams 1985; Diamond 2003). In a way, philosophy, more than other disciplines, is in a good position to ask where its limits might be – this is surely a philosophical question, not one that I will even try to answer, but hopefully one towards which this thesis might indirectly serve as a contribution.

The second interest is in literature. One could say art more generally, but literature in particular, and even more in particular, novels. I have always loved reading novels but, at the same time, I've always had an uneasy sense that I cannot fully see how they fit into my life. Of course, I didn't construe my experience as 'an uneasy sense that I cannot fully see how novels fit into my life' when I was in my teens, but even so I remember reading Dostoyevsky's *The Idiot* and asking my father what exactly we learn from such a novel. I didn't receive a satisfactory answer, nor did I come up with one. I felt a little stupid. It was only much later that I realised how hard it is to formulate something about what we learn from novels and that I was far from alone in my bafflement.

The question of literary cognitivism, as the question of what we can learn from literature is usually referred to, has also received significant attention over the past 40 years or so. Some attempts to give a cognitivist account, especially early on, have construed literature too much as a supplement to moral philosophy as done in the analytic tradition. Typical in this sense is Noël Carroll's (1998) 'clarificationism', which claims that literature can deepen our understanding of principles we already know by encouraging us to apply them to special cases. To use his example, white readers might have principles that are supposed to apply to all persons, to the effect that they treat them with respect, but through a work of fiction the fact that they are supposed to apply to African-Americans can strike home. I see where Carroll is coming from, and I don't think he's wrong, but if that were all that we can learn from literature, then so much for it. His answer in a way avoids the question, avoids wrestling with what great novels try to do. *Middlemarch* does contain the perspective of a woman whose depiction might have deepened some Victorians' understanding of their already formulated moral principles, but it is much more than than. More sophisticated and more influential, Martha Nussbaum's (1990) view provides a more ambitious answer. I will suggest that it is still too much tied to the moral philosophical way of thinking. I hope that the account I will give avoids this trap.

Of course, I will not try to give an account of *all* that we can learn from literature, nor do I think that it would be a wise idea for someone to attempt that. The question 'what can we learn from literature?' might be just as broad as 'what can we learn from philosophy?' or 'what can we learn from our friends?', so there could be mutually compatible theories of literary cognitivism. There are thus some views with which I will not engage, because they are not competing with the one I will put forward. Amongst them, I would mention those of Wayne Booth (1988), Frank Palmer (1992) and Jenefer Robinson (2005), which, along with Martha Nussbaum's, are some of the most interesting and original views of literary cognitivism that I have come across.

Although I am sympathetic towards recent attempts to narrow the perceived gap between fiction and non-fiction (Friend 2012; Matravers 2014), I will focus on fiction. Besides some fairly minor points, most of what I say should apply to biographies or other pieces of non-fiction as well. Also, I will focus on novels, and mostly on pre-modernist novels, which do not involve unreliable narrators, jumps in time or other techniques that in one way or another question the traditional novelistic form. I think this is methodologically sound, in that I will avoid complications to what I am going to say and make the thesis clearer, but this does not in any way rule out further applications of the views I shall propose to various narrative texts or indeed other art forms.

Enough with musing! We can now get to the plan of the thesis, which can be seen as consisting in two parts. In the first part (chapters 1-3), I will develop some views in the philosophy of mind, that I will then use in the second part (chapters 4-6) in order to give an account of what we can learn about attachments through literature. The first part is self-contained, and can be read independently of the second part, while the second part relies heavily on the first.

In the first chapter, I will give a theory of what attachments, that is, the individualised forms of caring such as those involved in love and friendship, are. The view that I will propose comes in two steps. The first one is that an attachment is *not* an emotional (or behavioural) disposition, but a mental state that underlies such a disposition. This state would manifest itself in emotions, but would not be constituted by them. The second step is elucidating what this mental state is. I argue against two dominant views: the 'properties view' (Badhwar 1987), which claims that this mental state is an appreciation or valuing of the traits of the person (or thing) one is attached to; and the 'relationship view' (Kolodny 2003), which claims that it is a form of valuing of the relationship, that is, roughly, the common history. The positive proposal I will put forward is based on a view of the mind according to which our emotions are the manifestations of what I call 'drives', that is, deeper mental states that encapsulate what we ultimately care about. The thesis that I will argue for is that an attachment just *is* such a drive, directed at the person one is attached to.

Given that we understand our attachments via the emotions we have, we should get as clear as possible on what emotions are, which is what I will do in the second chapter. I will defend a version of the attitudinalist theory of emotions of Julien Deonna and Fabrice Teroni (2012), which claims that an emotion is a bodily attitude directed at a content that is borrowed from another mental state. For instance, in an episode of fear, we first come across the object of fear in another mental state – for instance, we see it. Then, our body reacts, yet the physiological changes are directed at the content of this other mental state, with the phenomenology of the emotion dependent both on the content and on the bodily attitude. I will depart from Deonna and Teroni in arguing that the emotions that are manifestations of attachments do not have fittingness conditions. I will also argue that emotions directly motivate actions and, importantly, render courses of actions intelligible to the agent.

We have some kind of direct access to our own emotions, but in order to understand others, we need to imagine their emotions. This is what I will address in the third chapter, where I will provide a theory of imagining emotions. I will argue against the simulationist theory, which claims that imagining an emotion is simulating the emotion, that is, forming a mental state with a similar phenomenology, but different functional properties (e.g. it does not issue in behaviour). The positive view that I will put forward is based on a distinction inspired by Richard Wollheim (2015 [1980]): a representation is called 'thick' if the object is not merely referred to in that representation, but is apprehended in it, and 'thin' if it is not apprehended in it. A word, for instance, is a thin representation, while a drawing of an object is a thick one. The claim of this chapter will be that imagining an emotion is forming a thick meta-representation of that emotion.

Moving on to the second part of the thesis, I will first explore, in the fourth chapter, a theory that stems from a similar intuition to the one that is driving my thesis, the intuition that in reading literature we learn something by understanding fictional characters. This theory, known as the 'subjective knowledge theory', claims that in reading literature, we can get experiential (or subjective) knowledge, that is, knowledge of what some emotions (or other experiences) are like (Walsh 1969, Kajtár 2016, Bailey 2023). I will offer a defence of a version of this theory, while at the same time showing its limits in accounting for what we learn from literature. I will first argue that the experiential knowledge we get from the emotions we experience ourselves in response to fiction is not significant, as the range of emotions we can experience is limited. Instead, I will argue that the more promising version is that we gain experiential knowledge by imagining the experiences of characters. On the lines of David Lewis (1990), I will argue that the best way to construe experiential knowledge of an emotion is as an ability to imagine and recognise the emotion, from which it follows that the knowledge gained by imagining an emotion need not be worse than that gained by experiencing that emotion. I will end by observing that knowing what the emotions of a character, or of a person more generally, are like does not amount to understanding that person, for, as I will have argued in chapter 1, very similar emotions can be manifestations of very different drives, so the subjective knowledge theory does not quite account for the importance and cognitive rewards of understanding a character.

With these thoughts in mind, we move on to the fifth chapter, in which I will offer my positive view. A work of literature can acquaint us with an attachment of one literary character to another and in this way put us in a position to pass an ethical judgment on it. The work not only helps us imagine the emotions that the character has, emotions that reveal their attachments, but depicts the relevant emotions such that we can piece them together and understand the attachment. Of course, this is not a full acquaintance, in that there might be aspects of the attachment that one has not grasped, yet it is a good enough acquaintance, that puts us in a position to judge that attachment. I will illustrate the proposal by discussing Edith Wharton's *The age of innocence*, and then show that my view explains better the ethical ambition of this and other novels than the Aristotelian view of Martha Nussbaum (1990).

However, just being acquainted with an instance of an attachment and judging it does not seem to amount to a significant cognitive gain; the ethical knowledge that we gain seems to be about that instance and not about attachments in general. In the sixth and last chapter, I argue that in some cases it does amount to a significant cognitive gain. The key thesis of this chapter, inspired by Michael Tanner (2003), is the following: the value of an individual attachment is revealed in that instance and is only minimally connected to the concept of an attachment or to a broad description that applies to it. For example, the value of attachments involved in instances of friendship and love is not grasped by thinking in general about friendship and love and what aspects of them might make them valuable, but by understanding those instances. General judgments about attachments will turn out to be secondary to individual judgments and based on assemblages of individual judgments. We can now see what role literature plays: by acquainting us with remarkable instances of attachments, it shows how, and in which ways, attachments can be valuable.

CHAPTER 1. ATTACHMENTS

1. INTRODUCTION

Ethicists have come to accept more and more the importance of love, friendship and various other individualised forms of caring, to recognise that life is not only about responding to impersonal value but also, or even primarily, about the attachments we have. In this first chapter, I will build a theory of what these individualised attachments are, a theory that does not so much characterise various types of attachments as provide a framework in which we can understand what an attachment is and in which we can further discuss what individuates various types of attachments. My approach to this topic starts from the idea that we need a very careful discussion in philosophy of mind, a discussion that attempts to conceptualise what motivates an agent. It is for this that we will need the notion of a drive, that is, a fundamental standing mental state that manifests itself in emotions and behaviour and which captures the idea of 'what we really care about'. The thesis that I will then put forward is that an attachment just is such a drive, that X has an attachment to Y if and only if X has a special drive directed at Y that encapsulates Y's importance for her, a drive that I will try to characterise in this chapter.

In the second section, I will home in on the phenomenon of attachments, discussing various ways in which they have been approached in the literature, various intuitions that we have regardin them and various quandaries they provoke, throwing up many balls that I will then try to keep in the air for the rest of the chapter. In the third section will come the first major claim: that even though we associate the attachments with emotional and behavioural dispositions, the attachments do not consist in these dispositions, but are deeper mental states that manifest themselves in these dispositions. In the fourth section, I will discuss two popular ways to account for these deeper states, the properties view and the relationship view, and reject them. In the fifth section, I will present my own view, which will hopefully solve all the problems.

2. HOMING IN ON ATTACHMENTS

Attachments are woefully elusive, both conceptually and in practice. They are all around, yet their nature is very hard to encircle and point to. So an attempt to give an account of attachments should start with an an effort to locate the phenomenon better, as well as with some intuitions that will help us get going.

2.1. THE QUESTION

Let's start with an intuitive discussion of the phenomenon we are interested in. I mentioned love and friendship. When we refer to a relation like these, we think of emotions and actions that are directed at the other person. If X is friends with Y[1], X usually has various emotions towards Y, depending on the circumstances: joy when seeing Y, pride when Y achieves something, but also jealousy when she feels replaced by someone else in Y's life or grief when Y passes away. Besides feeling emotions, X might also perform various actions, such as helping Y when he needs it, having lunch with Y or sharing some ideas with him. Now, we can also have similar relations to groups of people, such as nations or institutions, or to inanimate objects such as buildings or works of art. One can have emotions towards a building and perform all sorts of actions regarding it, such as studying its architecture and history or conserving it. What is common to these relations is that they are relations to individual entities[2] and that the individuality of these entities is essential to them, which is what makes them individualised relations or attachments[3]. For the sake of simplicity, I will focus mostly on attachments

[1] For the sake of simplicity, I will use the feminine pronoun for X and the masculine for Y.

[2] It is open to debate what is and what isn't an entity, but this is beyond the point of the present paper. I will call an 'individual entity' anything that an agent represents as an individual entity.

[3] In the literature (e.g. LaFollette 1996) one often finds the term *personal relations*. I avoid using this term because it might imply that the relation is between persons, as

to other people, which will serve as examples, but the discussion can be easily extended to attachments to other kinds of entities.

To get a better grip on attachments, we can compare them to other kinds of relations. For instance, we can contrast them with universal love, which is supposed to be directed at any person solely in virtue of their being a person. Someone having universal love would have emotions towards all other people and perform actions in relation to them, but these would be manifestations of universal love and not of relations to individual people. Similarly, we might think of the professional relation that an employer might have with her employees: the employer feels proud of them and treats them with respect, but again, this is in virtue of a general concern for employees, as opposed to individual attachments to them. As the employees change, the employer *extends* the same relation to the new ones.

Now, importantly, I want to analyse these relations whether or not they are reciprocated. Of course, the English language has it that 'unrequited friendship' is a contradictory phrase, but this is a mere semantic problem – what I am interested in is what happens in the mind of someone that leads them to form a friendship, something that can happen whether or not the attachment is reciprocated. I am of course not claiming that reciprocation does not matter, that it doesn't change anything, just that even in perfectly symmetrical relationships we can analyse one side of the relationship, and that this is necessary in order to understand the whole. Given this way of framing the question, I call attachments 'relations' and *not* 'relationships', because the latter might be taken to imply some form of reciprocation. I will henceforth use the word 'relationship' to designate the series of emotions the two people have towards each other and behavioural interactions between the two, in short, their common history.

Next, and perhaps more controversially, I want to claim, again stretching the English language, that individualised attachments need not be positive, what might be imprecisely called instances of love. I will also include hate, as long as it is hate of an individual, in its individuality, and not hate of people with moustaches or of people that speak very loudly.

opposed to between a person and an entity, and thus excludes some of the cases I have just described.

There might also be attachments which cannot be easily classified as positive or negative.

It is already at this point that someone might frown and question the motivation behind the project. Why think that there's something interesting to be said about this large and inhomogeneous class of relations? Why focus on one side of the relationship of friendship, which we know has to be two-sided? Why think that love and hate are the same type of relation? Why not focus on something simpler, why not ask, like many philosophical papers and pop songs, 'what is love'?

The reason I ask precisely this question is, in short, that I have an answer to it. I think all these relations have something very important in common and that understanding what they have in common is essential to understanding each of them individually. Here, there is a methodological point that I want to stress. I don't think we should assume from the beginning that various phenomena are grouped together in a certain way, with the categories given to us by common language, categories such as friendship, romantic love, brotherly or sisterly love and so on, with our task being to 'give an account of' each of these. Rather, a philosophical theory should aim to illuminate, to show *why* we group phenomena in a certain way. So the reason for carving the world as I do should be clear not at the beginning but at the end. I am willing to give up the incipient 'cool, let's see' reaction from the reader and assume a bit of initial perplexity, with the hope of extracting a final 'aha'.

With the object of study somewhat delimited, let's proceed to some general intuitions about attachments.

2.2. LOVE *DE RE*, NON-FUNGIBILITY AND OFFICE LOVE

I will now discuss a series of related thoughts that pertain to the idea that attachments are towards individuals and not, for instance, towards people with certain characteristics. One way of phrasing this, following Robert Kraut (1986), is to say that attachments are *de re*, directed at individuals, and not directed at properties. To illuminate this distinction, let's see what it could imply.

For one, as we often do in philosophy, we might think of replacing a person with a duplicate. Suppose that an agent is friends with someone, and we replace that person with a duplicate that is qualitatively identical, physically and mentally, but numerically distinct. If the agent *really* is attached to the original person, the thought goes, they shouldn't 'continue' the old friendship with the duplicate. The friendship is with the individual, not with all its potential duplicates (cf. Kraut 1986, p. 421).

Second, as a variation on the first point, even though one formed a relation in some way in response to some properties of another person, say to their wit and lively eyes, one shouldn't then be willing to *trade up*, that is, to replace the friend with another person who scores higher on those traits, being wittier and having livelier eyes (Nozick 1989, p. 76). Something should have happened that bound the agent to the other person.

Of course, these intuitions are vague. If a friendship with someone ends and around that time a friendship with someone else commences, what would make this a form of replacing the previous person or of trading-up? Or if one's friend passes away and one begins a new, perhaps similar, friendship? Kraut (1986, pp. 428-9) seems to believe that this is something to some extent conventional, depending on the expectations of the participants in a relationship – e.g. 'if you marry someone else within three years of my death, you treat me as replaceable; four years is fine.' I think that only as a last resort, in case we find no revelatory account of what counts as replaceability and what does not, should we treat this as solely a matter of convention. For now, we can maintain hope of a more illuminating account.

As a third observation regarding attachments being *de re*, there is a sense in which X's attachment to Y does not always depend on Y's retaining all, or even most, of the characteristics that initially drew X to Y. There is a lovely passage from Robert Nozick which tries to capture this intuition:

> Perhaps we should think of love as like imprinting in ducks, where a duckling will attach itself to the first sizable moving object it sees in a certain time period and follow that as its mother. With people, perhaps characteristics set off the imprint

> of love, but then the person is loved in a way that is no longer based upon retaining those characteristics. (Nozick 1989, p. 75)

Of course, this all sounds very good, but we need to say what the imprinting consists in in order to really have an account of what an attachment is.

Another idea along similar lines to the ones discussed above is that of the undesirability of 'office love'. Kraut imagines Lisa, a girl, arguing that her mother does not love *her* because the mother hardly knows anything about what is important to her (1986, pp. 425-7). He couches this suspicion in terms of *office love*: what Lisa's mother loves is perhaps not so much Lisa as the office of her daughter. In some sense, the mother loves her daughter, whoever that might be, and not her actual daughter, Lisa. Now, Kraut is quick to dismiss this worry, claiming that nothing more than minimal knowledge is needed for love.

I agree with Kraut that extensive knowledge of the daughter is not required for the mother to count as loving her, but I think that the worry about office love is a real one. Lack of understanding of the daughter's personality might give rise to the suspicion of office love, but is not constitutive of it. Let's take another example, hopefully more revelatory. Suppose that X comes to think that it is good to have a friend and decides that she wants to have one. Knowledgeable as she is of human psychology, she looks for someone who shares her interests and has the resilience that might conduce to long-term interactions. Out of all her acquaintances, Y seems to fit best. She thus decides to become friends with Y: she acts in a friendly manner towards Y, engages Y in discussions about their interests and forces herself to get concerned about Y's projects. Keeping in mind that X really cares about her goal of having a friend, it is plausible that she is content when she seems to be doing well in achieving this goal – for instance, she feels happy when she schedules an activity together with Y. One day, though, Y comes along and says to her: 'You don't truly love *me*. You just wanted to have a friend and picked me to fill the role.' Note that even though this might look like a charge of fungibility, it doesn't imply that X would be willing to exchange Y for someone who filled the role better. Y might be perfectly aware that if X is principled enough, or obstinate enough, she might very well maintain her intercourse with Y for all her life and avoid having any other similar relationship. Rather, the charge is that in the creation of the relationship, Y didn't count enough, or not in the right way. Y's thought is that X merely created an office when deciding to have a friend and that the creation of the office was in a way the creation of the relationship, a creation to which Y had no relevance.

Someone might of course object that even though Y did not count in the right way in the beginning of the relation, this does not preclude a genuine attachment having been formed in the meantime. I totally agree with that, but my point is that it need not have formed. So the worry about office love is that a relation might involve positive emotions, care, spending time together *and* some form of non-fungibility without thereby amounting to a genuine attachment.

2.3. NO THOUGHT TOO MANY

The second idea about attachments stems from Bernard Williams' famous one-thought-too-many discussion[4]. The thought experiment discussed is as follows: a man (who is informally assumed to be Williams himself, so I shall stick to the convention) is faced with two people in peril of drowning, his wife and a stranger. Only having time to save one of them, he has to make a decision. The deontologist (and the utilitarian) should find it easy to judge the case: they would concede that Williams should reason that if he can only save one person, the fact that one of them is his wife is a legitimate reason to save her. His thought should be something like 'she is my wife and in a situation like this it is permissible to save one's wife', or even if he doesn't consciously entertain the thought, his motivation should be along these lines. Williams' contention is that this is exactly what shouldn't happen, that his motivation should be simply 'she is my wife', and that any further thought is 'one thought too many'. Essentially, what Williams seems to be getting at is that a true attachment to his wife would give rise to motivations that are not filtered through the algorithms of impartial

[4] Williams (1989, pp. 18-9). The thought experiment Williams discusses is actually Charles Fried's.

morality, motivations that lie outside the impartial morality. Frankfurt (2004, pp. 36-37) argues that even the thought ‘she is my wife’ is one thought too many, that the full motivation should be something like the unedifying ‘she is she’. The thought ‘she is my wife’ would point to some general commitment to save one’s wife, whoever that may be. Frankfurt is a bit pedantic, as this is probably what Williams meant, but this pedantry reveals the fact that it is not straightforward to characterise the *right* motivation (assuming that we agree with Williams’ intuition).

Troy Jollimore (2011) proposes a solution. He starts from an Aristotelian framework in which a central element in decision-making is *perception*, that is, the quasi-automatic recognition by the agent of factors that are relevant to the decision to be made. The virtuous agent should not take into account every single aspect of a situation and ponder what relevance it might have. Rather, they are supposed to immediately *see* the aspects that are relevant and essentially construe the situation in terms of these forms of relevance. Jollimore thinks this framework could help us understand what is wrong in the one-thought-too-many case and how the ideal husband would act. Not only would he save his wife, but in perceiving the situation, he would instantly see the peril to his wife as a reason and as *the only relevant reason*. In Jollimore’s words, the husband is not only to see the peril to his wife as a consideration, but

> it is also necessary that he perceive this consideration as possessing such overwhelming importance that *it simply drives everything else from his mind*. If the danger to [his wife] does not strike him with the sort of force and practical import, then his love for her is shallow or not entirely genuine. (Jollimore 2011, p. 35, his emphasis)

Perhaps Jollimore is right that this is how the husband should perceive the situation, but I think he does not get to the bottom of things. Indeed, we might compare this case with the following: suppose one has to choose between saving a stranger’s life and giving another stranger directions in the street. Of course, one should save the stranger and, moreover, one should perceive the situation in such a way that giving directions does not even enter the mind as a possibility. But this is just because one consideration is way more important than the other, not because there is any attachment involved. In the case of the drowning wife, I think that Williams' point is not so much that his wife being saved should be vastly more important to him than that the stranger should be saved. Rather, it is that it is a *different kind* of consideration. So we still have to explain what the difference is between seeing one consideration as vastly more important than other and seeing one consideration in virtue of an attachment.

Furthermore, we can think of a variation on Williams' case. Perhaps the choice is between saving one's wife and saving many other people at once (a hundred, a thousand). It is of course difficult to say what one might do in such a case, but I take it that Jollimore would concede that the love of one's wife need not drive the thought of all the other people out of one's mind. Yet, we might still want to say that, regardless of the husband ends up doing, if he has an attachment to his wife, the consideration to save her is of a different kind than the consideration to save all the other people.

We get a similar worry from Michael Stocker (1976). Stocker asks us to imagine a man visiting his friend in hospital. When the friend thanks him, he just says, *and means*, that he only came because it is his duty to visit his friend in the hospital. We can then imagine what the friend feels: having been so happy that the man thought of him and wanted to cheer him up, he realises that it has nothing to do with *him*, but only with a general sense of duty. Again, the fact that the man, as a good Aristotelian agent, was so taken by the impetus to visit his friend in the hospital that all other considerations were silenced would do frightfully little to alleviate the disappointment of his friend.

What Williams and Stocker are driving at is that an action filtered through some kind of impartial algorithm doesn't reflect a true commitment to a loved one, or in my language, a genuine attachment. But they don't really spell out what kind of motivation *would* reflect such a relation. It shouldn't stem from a general principle, but this is clearly not enough. One can just perform actions on a whim, and presumably these do not reflect an attachment. Or one could save a stranger instinctively, without basing one's decision on some impartial concern, and that wouldn't reflect an attachment either, for the simple reason that there is no attachment to that stranger in particular. In a later paper, Stocker suggests that an action displaying the attachment would somehow be 'out of the friendship' (1981, p. 748). This is begging the question: if we want to identify the actions that belong to a friendship, it is not very useful to say that they are those that are done 'out of the friendship'. Yet even though it begs the question, the phrase 'out the friendship' sounds right, and I hope to show that this intuition is correct.

2.4. STATIC AND DYNAMIC ATTACHMENTS

Discussions of love *de re*, irreplaceability and the like might create the following caricature: that if X has an attachment to Y, X loves Y in the same way, irrespective of how Y changes, X hopes that Y's plan are fulfilled, irrespective of what they are, and so on, in short, that X is like a loving machine. The worry is that even if X's attachment to Y is *de re* or whatever, it does not meaningfully take into account anything that matters to Y. In other words, it doesn't matter to X whether Y cares about something or something else. Amélie Rorty has this sort of worry in her exchange with Kraut and proposes instead a different ideal of attachment:

> There is a kind of love – and for some it may be the only kind that qualifies as true love – that is historical precisely because it does not (oh so wonderfully) rigidly designate its objects. The details of such love change with every change in the lover and the friend. Such a love might be called *dynamically permeable*. It is *permeable* in that the lover is changed by loving and changed by truthful perception of the friend. […] Having been transformed by loving, the lover perceives the friend in a new way and loves in a new way. (Rorty 1986, p. 402)[5]

[5] Similar thoughts are found in Cocking and Kennett (1998) and Nehamas (2016). Cf. Delaney (1996) for an alternative approach to how romantic attachments should change.

What does Rorty suggest when she says that the lover should be 'transformed by the loving'? It seems that there are essentially two kinds of changes that might happen (cf. Cocking and Kennett 1998). The first is a change in the attachment. New events and new discoveries about Y might change the way X relates to Y. For instance, a discovery that they have very different opinions about music might lead to a kind of friendly rivalry, in which X teases Y when his favourite singer performs badly – 'And you still admire her after that concert? Wow, you really are a diehard fan.' The second kind of change is in aspects of X's life that are not related to Y. Perhaps seeing Y playing the violin causes X to become interested in music written for the violin, an interest that might survive Y's loss of interest in the violin or even the end of their relationship. Even if caused by her attachment to Y, her interest in violin music has now a life of its own, and if the relationship ends somehow, then, as Rorty suggests, X's interest in the violin is a testimony to that relation.

It would then follow that there is a 'dynamism' that is valuable in certain attachments, that both the person having the attachment and the attachment itself change in light of the other person and of their interaction. These attachments might be contrasted with those in which the person and the attachment do not change so much, and which can thus be naturally called 'static'. Note that it is not necessary that the more dynamic an attachment is, the better. Perhaps there is something to be said for the love of one's children being static. It is often what they take the paradigm case of love to be that leads philosophers to build various theories: Rorty seems to take the paradigm case to be that between two adults, while for instance Frankfurt explicitly takes it to be that of a parent to her child (2004, p. 43).

A theory of attachments should account both for more static and for more dynamic attachments and explain whence the difference stems. Rorty seems to think that there is a tension between an attachment being *de re*, which makes it intuitively in some way unconditional, and its being dynamic. I aim to explain away this apparent tension, to show that there can be attachments that are both *de re* and dynamic.

3. ATTACHMENTS ARE CATEGORICAL, NOT DISPOSITIONAL MENTAL STATES

I will now take the first step in explaining what attachments are, by arguing that they are *not* mental dispositions, but categorical mental states that underlie mental dispositions. In the next sections, I will give an account of what the mental state that underlies the disposition is.

We get an idea of what attachments might be from their manifestations. Indeed, when we say that X seems to have an attachment to Y, we get this impression from X having positive emotions towards Y, her engaging in caring behaviour towards Y and so on, and from her doing this in a more or less regular fashion. Of course, the way in which we come to get an idea of a phenomenon need not hand us directly the nature of the phenomenon – we get the idea of heat by touching hot things, but the sensation that we get is somewhat far, conceptually speaking, from the nature of heat – yet it is a natural assumption that the attachment just *is* a disposition. Such a disposition would be a disposition to have emotions, such as hope for the other's plans, gaiety when one sees the other, etc. It might also include behavioural or motivational dispositions, such as the disposition to take the other's happiness as a motive in one's decisions. I'm using the phrase 'might include' regarding behavioural dispositions because it could be that the motivation for action stems directly from the emotions; in that case, we don't need to posit a behavioural or motivational disposition in addition to the emotional one. These are details to be filled in.

I will argue that this conjecture is incorrect, that the attachment is not identical with the disposition. The account of attachments as dispositions depends, of course, on what we take dispositions to be, hence this discussion cannot be kept separate from discussions of the metaphysics of dispositions, but I hope that I can make my case without going too deeply into the latter. I will consider two ways of cashing out the mental dispositions that an attachment is supposed to consist in: according to the first version of the theory, the disposition would just be a set of hypothetical conditional statements; according to the second, it would be a genuine property of the mind.

3.1. ATTACHMENTS AS SETS OF CONDITIONALS

According to the first version of the theory, X has an attachment to Y if and only if a set of hypothetical conditional statements are true: that she tends to engage in caring behaviour towards Y if he needs it; that she has hope of success if Y is working hard for a goal; etc.[6]

Of course, it is very hard to come up with a set of conditionals that would characterise all attachments, but we can think of various ways around this problem. Perhaps we can come up with many conditionals such that if a relevant proportion of them is true, X has an attachment towards Y. Or perhaps attachments come in degrees, so that X is more attached to Y if more conditionals are true. Or, as a third version, there are many sets of conditional statements, each corresponding to a type of attachment, and then X has an attachment to Y if and only if one set of conditionals is true.

The first problem with this view is that according to it attachments are rather flimsy. The conditional analysis seems to imply that discovering what attachments are involves little more than conceptual analysis, and little philosophy of mind or ontology. Furthermore, it is rather unsatisfying that this account would not quite explain what it is that unifies all the emotions in the dispositions, why we take them to be part of *one* phenomenon, i.e. the attachment. We might hope that attachments are more part of the fabric of the world in a way that explains what it is that binds all these emotions together. Of course, if no convincing account can render them thus, we might rest content with the view at hand, but we should strive for a better one, with more explanatory power.

The second problem with this view is that it cannot fully account for the difference between X having an attachment to Y, and X having a disposition towards Y in virtue of another attachment or commitment. For example, suppose that X has a very strong attachment to her school. This leads her to have positive feelings towards its alumni, of which Y is one. In

[6] For such a view of dispositions in general, see, for instance, Goodman (1954).

virtue of this, X has positive feelings towards Y and is disposed to care about Y. The conditional statements that define whether X has an attachment to Y – statements like ‘X would be happy if Y succeeds in his career’ – would be true, and this would license us to say that X has an attachment to Y. Yet we want to say that if X has very similar dispositions towards all alumni of the school, she doesn’t have an attachment to Y in particular.

The third problem is related to the second one. Suppose that X has an attachment to Y, but also an attachment to her school, of which Y is an alumnus. She will then have all sorts of positive emotions towards Y in virtue of her attachment to Y, but some of her positive emotions towards Y might be at least partly manifestations of her attachment to her school. For instance, suppose that X’s school has a tradition of producing good lawyers. In that case, X’s hope that Y succeeds at the Bar exam might be related to the tradition of the school and in this way manifest X’s attachment to the school rather than her attachment to Y, or at least more than her attachment to Y. We thus need a principled way to explain how not all positive emotions of X towards Y are part of the attachment to Y, and the view under consideration does not seem to be able to account for this.

One way to solve these problems would be to claim that dispositions, in particular the dispositions that we are interested in, amount to something more than a set of hypothetical statements being true. This is the view that I will consider now, a sort of ‘realism’ about dispositions. Most of the objections that I will put forward towards this version will apply to the conditional version as well.

3.2. ATTACHMENTS AS ‘REAL’ DISPOSITIONS

The central idea of this second version of the dispositionalist view is that dispositions are not merely sets of conditionals, but rather genuine properties of the entities that have those dispositions. To give an account of dispositions then, one should not do conceptual analysis, but ontology (Molnar 2003). In the mental case, the dispositions would be genuine properties of the mind (or of the brain). One could then claim, as Hichem Naar does of love (2013, 2018, forthcoming), that an attachment would be just one such dispositional property. What I called ‘attachment’ is a broader concept than ‘love’, but we can easily extend Naar’s account to attachments.

This move would solve the three problems highlighted above. First, the attachments would be genuine properties, a fact which should alleviate the worry that, according to a dispositionalist view, they are flimsy.

Second, if X has the same kind of emotional reactions towards all people that went to the same school with her, Y being one of them, there would be only one dispositional property, that includes all of X’s emotions towards people from the same school. There would simply not exist another dispositional property, one that would account just for her emotions towards Y, so X would not count as having an attachment to Y. It would be a bit like a fragile globe which, besides its general fragility, does not have some extra properties of left-fragility and right-fragility that would account for its being disposed to break when thrown on its left and right sides, respectively.

Third, we can also return to the case in which X has an attachment both to her school and to Y, who happened to have gone to the same school, a case in which we want to give an account of how to distinguish an emotion she has towards Y in virtue of the attachment to the school from one that she has towards Y in virtue of her attachment towards Y. There could be a few ways to explain the difference, and one is the following: it is sometimes assumed that if dispositions are genuine properties, they have a causal impact on the world (Mumford 1998, pp. 118-43). In our case, then, we can say that the emotion is a manifestation of one of the attachments insofar as it is caused by that attachment, either directly or in conjunction with other dispositional properties. So, whichever disposition is the one that caused X’s emotional reaction, it is to that attachment that we should ascribe the emotion. If it is caused by both dispositions, we should ascribe it to both attachments.

Even though this account solves these problems, I think it is ultimately unsatisfying. Although Naar does not explicitly say this, I take it that he accepts that the emotional dispositions towards other people have some categorical basis, that is, that there is a categorical property in virtue of which X has the dispositional property that constitutes her attachment to Y. So, the question is, why not identify the attachment with the categorical property? What we are interested in is to understand the agent, and it seems plausible that categorical properties capture better this phenomenon than the dispositions that are manifestations of these categorical properties[7].

But I will bring more substantial reasons to identify the attachment with a categorical mental state rather than with a dispositional one.

A key worry is that the dispositional account ties the attachment too much to the manifestations. I will discuss three scenarios which will show that this approach is wrong-headed.

In the first case, suppose that a father has a child to whom he has a strong attachment, with the effect that many of his emotional reactions are directed at his child, and a lot of his behaviour is related to his child. For instance, when he sees some sweets in a shop, he instinctively wonders whether his child would like them. He then fathers two more children, and assuming he loves them similarly, his attention and emotions are divided among all three, and hence the first-born receives less attention than she used to. So the father's disposition towards the first born has changed, yet we want to say, I think, that his attachment has not changed. This would imply that the attachment is distinct from the disposition. Something clearly has changed, but something essential has remained the same, and I don't think that the dispositional theory can account for this.

In the second case, we have again our father, this time not having any further children, but getting depressed, to the effect that he stops having his usual feelings for his family (Goldie 2011b, pp. 99-100). After one year, the depression goes away and his disposition towards his family returns exactly as before. Naar (2013) uses this example to support his realist dispositional account as a better option than the conditional analysis examined in the previous subsection. He starts from the assumption, which I am happy to accept for the time being and will argue for later, that the attachment remains throughout the depression. Then he explains how this is

[7] According to Mumford (1998, pp. 144-69), a dispositional property and its categorical basis are the same 'property instance'. The dispositional aspect and the categorical aspect are just different ways of talking about the same thing. If this were the case (though I struggle to make sense of the theory), I suppose I would be happy with accepting that the attachment just is that 'property instance'. However, as will become obvious, I think that a person can keep an attachment unchanged, while the disposition that that attachment manifests changes. I am not sure how we can make sense of this on Mumford's view.

a form of *masking* of the disposition, a bit like fragility can be masked by wrapping the object in protective foam[8]. He claims that only a realist view of dispositions can accommodate the phenomenon of masking, as there really is a property to be masked, whereas in the conditional analysis, there is no property to be masked.

It is not perfectly clear what this 'masking' amounts to, so we could usefully unpack the metaphor. One possibility is that the depressed father feels and behaves as if he had no attachment whatsoever to his family, essentially as if they were strangers that he couldn't care less about. This is a possibility, but it is a far-fetched one, and if the father really exhibited this sort of extreme indifference, one might wonder whether he had really kept his attachments to his family. A more plausible description, and one that covers more cases, is that the father's emotions and behaviour are changed by the depression. He might be indifferent in some situations, but in others he might be, for instance, more blunt than he would have been before the depression settled in: if his child does something wrong, he might explain to him why it was wrong without the usual care and softness. Or, if his child achieves something and everyone around him is happy, he might feel some regret that he cannot be as happy as the event is significant for his child. All these reactions, different as they are from his usual ones, are still manifestations of the attachment. Therefore, the masking metaphor is a bit misleading.

To explain better what is going on, I think we should distinguish two cases. One option is that the depression might be the result of something bad having happened. Perhaps a very good friend of his has passed away. In such a case, I think that the phenomenon is similar to the case in which the father has two more children. The emotions that the father has are the result of more than one attachment. When we think of the emotions towards his family, the attachments to his family are at work, but the attachment to his friend is at work as well, and because his friend has passed away, there is a sense in which the attachment to his friend has become more influential.

Another possibility is that the depression does not have an explanation that appeals to another attachment or another important

[8] Here, he takes the idea from Johnston (1992).

commitment – one just is depressed. In that case, it is as if there is some kind of external influence on the emotions and behaviour of the father, an influence that he cannot make sense of from the inside.

In both cases, what happens is that the emotions of the father are the result of his attachment to his family *and* something else, whether of another attachment or of an external cause. This is a more precise and revealing way of explaining the phenomenon than saying that the attachment is somehow 'masked' by something else. Anyway, in both cases, we see that depending on whether or not the father is depressed, the manifestations of his attachments differ, or in other words, the emotional dispositions that he has towards members of his family differ. But, if we want to say, as Naar also does, that his attachments remain the same, we obtain again the result that the attachments are not identical to some emotional dispositions. Rather, they seem to be something that underlies emotional dispositions. Depending on various factors, the same attachments might underlie different dispositions.

The third case that I want to discuss is that in which our father experiences a change of values. For instance, suppose that his child play-acts with her friends, and the father enjoys seeing her. Then, he reads Austen's *Mansfield Park*, a novel that displays a negative view towards amateur theatricals and tries to show that these can be in various ways vicious and dangerous for human relationships. As a consequence, the father becomes convinced that there is something wrong with amateur theatre. He stops feeling joy when seeing his child play-acting with friends and starts feeling a kind of anguish. Yet again, there is a sense in which his attachment to his child remains the same and the dispositional view cannot account for that. It might be replied that this is a far-fetched case, but I'm not sure this is so – any feeling of joy that the father might experience towards his child depends not only on the attachment to his child, but on other cares and concerns that he has, so there is hardly any manifestation that the father has towards his child irrespective of other cares and concerns he has.

One move the dispositionalist might make against this third case is to try to build the cares and concerns into the definition of the disposition: 'The father feels joy when the child does something he cares about or finds worthwhile.' The problem with this move is that if we want to go the dispositionalist way, cares and concerns should also be defined dispositionally – e.g. to care about X and to find it worthwhile would involve, among other things, feeling joy when one's child does X. So we have a circularity. The way out that I am proposing is to see both the attitude towards amateur theatre and attachments in some categorical way, such that the dispositions that the father has are the result of many such categorical mental states that he has.

Having argued that we shouldn't tie the attachments too much to their manifestations, I want to make one last point against the dispositionalist theory of attachments. The argument is as follows. Two dispositions are similar insofar as their manifestations are similar, i.e. the similarity between dispositions should be cashed out in qualitative terms. If two vases are disposed to break in almost the same circumstances, their respective properties of fragility – their 'fragilities', one might say – are similar. Now, I want to claim that it is *not* true in the case of attachments that if two attachments have very similar manifestations, the two attachments are themselves similar. This would imply that attachments are not identical to the dispositions they lead to. To do this, consider the dispositions of two mothers directed at their daughters:

> *Mother A:* Mother A is highly involved in her daughter's ballet lessons. They both hope that she becomes a good ballerina and devote a lot of time to this. The mother accompanies the daughter to all the shows, takes her to the doctor to preempt any possible medical problems and feels joy when the daughter progresses. Sadly, when the daughter turns 16, her body develops such that it is impossible for her to become a world-class ballerina. The mother suffers with the daughter, yet encourages her: 'Now you can pursue your interest in philosophy, which is decent as an alternative.'
>
> *Mother B:* Mother B, exactly as mother A, is very involved in her daughter's ballet lessons. They both hope that she becomes a

> good ballerina and devote a lot of time to this. The mother accompanies the daughter to all the shows, takes her to the doctor to preempt any possible problems and feels joy when the daughter progresses. Sadly, when the daughter turns 16, her body develops such that it is impossible for her to become a world-class ballerina. The mother gets very annoyed and says to her daughter: 'I've wasted my time with you.' (and really feels that, not only in that moment).

Even if the emotions of mothers A and B are very similar, the only major difference being the final response, I take it that it is precisely the final response that reveals that their attachments to their daughters are very different. If we try to put our finger on the difference, the most natural thing to say would be that while mother A has a normal loving attachment to her daughter, which makes them build projects together, such as that of the daughter becoming a ballerina, mother B has a controlling attachment to her daughter, more like that of ruthless creator to her creation. It follows that even though the emotional dispositions associated with the attachments of mothers A and B for their daughters are similar, their attachments are very different. Therefore, their attachments are distinct from their emotional dispositions.

Now, if we assume that attachments are categorical properties that manifest themselves in dispositions, we can explain the phenomenon of the two mothers very easily. Indeed, nothing prevents two very different categorical properties from leading to very similar dispositions. To return to the example of a fragile vase, two vases of very different materials or forms can be disposed to break in very similar circumstances.

To conclude, even the second version of the dispositionalist theory faces significant worries: first, it is not clear why, if there is an underlying categorical property, we shouldn't identify the attachment with that rather than with the disposition; second, it ties the attachment too much to its manifestations and so cannot account for the fact that the manifestations can change while the attachment remains the same; third, it cannot account for the fact that two attachments can have very similar manifestations while being very different. These problems suggest that we should identify the attachment with a deeper, categorical mental state, yet in order to put forward a convincing account, we need to find out what this mental state might be.

4. TWO MAJOR VIEWS

Now, let's see what this deeper mental state that manifest itself in an emotional disposition could be. Before presenting my own view, it would be good to discuss some views existent in the literature. I take it that the dominant views are what might be called the *properties view*, which cashes out an attachment as a form of valuing the other's properties (Badhwar 1987, Keller 2000), and the *relationship view*, which cashes it out as a form of valuing the relationship that one has had to the other person (Kolodny 2003). These theories are not formulated in terms of categorical mental states, but I think that we can very easily reformulate them accordingly, without losing much of what their proponents aimed for.

4.1. THE PROPERTIES VIEW

When we form attachments, we often respond to the characteristics of the person (or entity) we form an attachment to. Indeed, we are often taken by the wit, looks, interests or way of seeing the world of the other. Of course, we don't see the other as a sum of characteristics, but experience them together, seeing their intelligence in their eyes, their character in their smile and so on, yet the fact remains that we respond to their characteristics. It would then be natural to conjecture that the disposition we associate with the attachment is the manifestation of a mental state of valuing certain kinds of characteristics. This is the characteristics view, or the *properties view*[9]. For the purpose of this view, I will assume that these characteristics are 'not overly extrinsic' (Keller 2000, p. 165)[10].

[9] This view arguably goes back at least to Aristotle (2000), and is defended by Badhwar (1987) and Keller (2000). Cf. Taylor (1976) and Delaney (1996).

Let's think a bit more about what is going on in the mind of the agent forming the attachment. On one interpretation, we have a system of valuation that we start from, that is, we value certain things independently of the relations we have, whether they be general characteristics, such as wit, intelligence or an interest in impressionism, or fairly detailed ones, such as intelligent-eyes-that-seem-to-pierce-your-defence-while-leaving-you-intact[11]. Then, when we encounter a person that instantiates to a very high degree the characteristics we value, our system of valuation gives rise to our emotional disposition towards that person, forming an attachment to them, of one kind or another. At the risk of being unfair to this account and making a caricature out of it, it seems like our valuation system takes as input characteristics, evaluates them, and provides as output emotional and behavioural dispositions.

This account is subject to the fungibility objection: if X has an attachment to Y solely in virtue of appreciating Y's properties, there might appear on the stage another person, call him Z, that instantiates the valued properties to a higher degree, that in Robert Nozick's words, has a 'higher score' on these characteristics (1989, p. 76). In this case, X's valuation system would divert her emotions and behaviour towards Z, and if there is some constraint on the number of relationships X can have, either in virtue of time or attention, or in virtue of the attachment to Y being a romantic one, X will essentially replace Y with Z.

There is an objection to the way I have articulated the scenario. I said that Z 'scores higher' than Y on various traits such as wit and intelligence. This suggests that one's wit is a quantifiable property, that one has a certain quantity of wit, which is to be measured in metric or imperial units. Yet one might claim, as Neera Badhwar does, that to think of a trait

[10] This is, of course, a rather intuitive description that avoids going into the hard debate regarding the difference between intrinsic and extrinsic properties. Being 'the first person I saw after graduating from university' is clearly an extrinsic property, while having 'brown hair' is intuitively taken to be, for the purpose of human interactions, a 'not overly extrinsic' property, irrespective of our view of the metaphysics of colours. The 'relationship view', which I will analyse in the next subsection, might be interpreted as relying on more extrinsic properties.

[11] I use the word 'value' because it is commonly used in the literature, even though I think it is very vague: it could refer to a judgment about objective value, a psychological state of appreciation with no claim to objectivity, or something in between. I don't think it matters how we cash out its meaning in the discussion that follows.

such as wit as a scale on which people score higher or lower is wrong (1987, pp. 19-23). The wit of different people is not commensurable on a scale, and what X appreciates in Y and Z is not the same quantifiable property, which Z has more of than Y, but rather Y's particular wit and Z's particular wit. If that is the case, one might hope that this alleviates the fungibility worry.

I agree that different people's wit does not consist in variable quantities of a uniform product, namely wit, but I don't agree with Badhwar that this alleviates the fungibility worry. Even though the traits of Y and Z are not on one scale, if X decides whom to have a relationship with based on appreciating the traits of Y and Z and valuing Z's traits more than Y's, X essentially *puts these traits on a scale*, according to how much she appreciates them. To see better that this is a fungible way of seeing other people, think of the following scenario: I want to buy a tasty sandwich. For this purpose, I compare the 'tastinesses' of various sandwiches, and even if these tastinesses are not straightforwardly instances of a quantifiable property of tastiness, as they are actually different kinds of tastinesses, I should still be able to rank them from the most to the least tasty and pick the tastiest. Morever, I can do this irrespective of whether I take myself to rank their 'objective tastiness' or just to order them according to the pleasure they afford my taste buds. It follows that even if their tastinesses are not straightforwardly instances of a quantifiable property of tastiness, for the purpose of my buying a sandwich, what matters is only where they lie in my ranking. Similarly, for X's purpose of having a friend, what matters is where other people's wit lies in her ranking. If Z's wit lies higher than her friend Y's, she would replace Y with Z[12].

Furthermore, besides the fungibility worry, it just seems fairly obvious that attachments are not always based on valuing the characteristics of the other person or entity. The most natural counter-example is that of

[12] The same worries apply to another attempt to rescue the properties view, based on 'perspectival properties'. Sara Protasi (2014) argues that we shouldn't see the properties that we appreciate in the people we love as properties that are accessible only from our perspective, based on our interaction with them. These properties represent how they appear to us. This would ensure that the properties that we appreciate in them are not merely instances of some quantifiable property. Yet again, the fact that we rank the properties of the others, even if perspectival, and feel emotions and act towards them based on where they are in the ranking, means that we treat them as fungible.

attachment to one's children. Another counter-example is that of people remaining friends or remaining together in spite of their losing some of the characteristics that initially drew one to the other. Of course, my opponent could claim that perhaps there are two different phenomena: there is one in which we form a relation to someone else based on our appreciation of them and another in which people stick together for a completely different motivation. We shouldn't exclude this possibility from the outset, but I hope that in the end my account of attachments will unify the two phenomena.

4.2. THE RELATIONSHIP VIEW

I will now proceed to what seems to be a more promising view, the relationship view (Kolodny 2003, 2010). The idea behind this view is that the motivation for an attitude of love, or more generally what I've called an attachment, hinges on the historical *relationship* one has had to the other person[13].

So, let's start by saying briefly what a relationship that X had to Y would consist in. It would consist in the emotions and concern they *have had* towards each other (i.e. in the past), all the interactions that they've had and that have manifested these emotions and concern, in short, their common history. It would also include facts about them being related (e.g. cousins) or part of a larger group or institution (e.g. colleagues). Moreover, from the perspective of X, one should add in the mix Y's current attitude towards her, that is, Y's present emotions and concern.

Now, Kolodny claims that it is a form of valuation of this relationship that should ground love. Indeed, he argues that:

> Love is not only rendered *normatively appropriate* by the presence of a relationship. Love, moreover, partly consists in the belief that some relationship renders it appropriate and the emotions and motivations of love are causally sustained by this belief (except in pathological cases). Special concern for a

[13] Note again, I am using the word 'relationship', not 'relation', to denote the history of the emotions and interactions.

> person is not love at all when there is no belief that a relationship renders it appropriate. (Kolodny 2003, p. 146)

What happens in X's mind, according to this view, is that she has a kind of valuation that sees a type of relationship R as rendering an emotional disposition appropriate and believes that she has an instance r of R with Y. So, even though Kolodny might not phrase things in this way, in the framework I have argued for above, the mental state that manifests itself in the disposition towards Y is a valuation of the type of relation R[14].

Of course, this view escapes the fungibility problem that the properties view had. It is clear that Y cannot be replaced with any other person Z, as there is no person Z to which X had the same relationship.

Now, the view has an immediate problem that Kolodny himself recognises and tries to remedy, namely the worry of how a relationship begins. For it seems that a relationship should be underscored by mutual attachment and, on the view at hand, mutual attachment is dependent on the pre-existence of a relationship. It would follow that a relationship cannot begin, as for it to begin, we need mutual attachment, which in turn depends on the relationship being already in place. Kolodny's answer is the following: a relationship starts with mutual liking, with doing things together, enjoying each other's company and so on, but *without* anything like what I've called attachment or individualised concern. Then, once the two people have developed a common history of this kind, they have a reason to develop an attachment or love for each other, to have an individualised concern of the kind that I've described. Yet according to Kolodny, the common history at this point amounts to a reason but *not to an insistent reason*. By this, he means that developing an attachment would be appropriate, but *not developing* one would also be appropriate. However, once mutual attachments are developed and they have manifested themselves in mutual interaction, the common history of the relationship

[14] Of course, X might *wrongly* believe that she has the right type of relation with Y. In that case, Kolodny's view would imply, I take it, that the problem is with forming the belief, not in the belief leading to emotions and behaviour.

becomes an insistent reason for the attachments to continue[15] or, in other words, not continuing the emotional disposition would be inappropriate.

There is another related problem that Kolodny doesn't address, but that can be addressed in a similar way, the problem of how the attachment can become stronger. Of course, once there is a relationship in place, according to Kolodny, one has a reason to continue having the attachment. But how does the attachment grow stronger? I think Kolodny would give a response as the one above: that at various moments, there is an insistent reason to continue the attachment as it was and a non-insistent one for it to become stronger.

Now, I shall present my case against the relationship view. First, there is a large bullet to bite in accepting that one cannot have an attachment towards someone whom one hasn't had enough interaction with and who does not have an attachment to one (e.g. an unrequited attachment). It might be, though I remain neutral on this, that one cannot have an attachment to someone one barely knows. But even if this is true, we can still have a scenario in which X knows enough about Y without having had the kind of relationship that justifies the attachment. According to Kolodny, it would be impossible for X to have at that moment an attachment to Y.

Kolodny does attempt to address the issue of unrequited love:

> If my concern for Lisa is fated to be unrequited, then it is open to a familiar kind of criticism, which may come first in the gentler form of advice to 'get over it and move on,' and later in the more forbidding form of a restraining order… [Our tendency to valorise unrequited love] does not reflect a conviction that unrequited love, as such, is somehow worthwhile. Although one feels a wet blanket for saying so, it is a simple fact that we do not encourage our friends in their futile pining in the way in which we might encourage them in their creative ambitions or actual relationships. Indeed, if it persists, we are apt to find it unsettling. Either our friends are in the grip of emotions that

[15] Of course, if nothing unusual happens, e.g. one of the two transforming into a monster, which might amount to a reason that outweighs any reason given by the common history.

> they themselves can no longer see the point of, or they have lost touch with the reality of their situation. (Kolodny 2003, p. 171)

In his subtle-but-not-so-subtle allusions to 'restraining order' and 'los[ing] touch with the reality of their situations', Kolodny conflates the attachments we have with a type of behaviour and portrays people loving unrequitedly as little less than stalkers. He obscures the fact that many people in such a position realise that the best thing to do in light of their attachments is just to leave the other person alone. Leaving the other person alone does not amount to the attachment disappearing and, even more to my point, the decision to leave the other person alone might actually *stem* from the attachment and the concern it involves. Moreover, Kolodny's discussion of whether unrequited love is or isn't worthwhile obfuscates the fact that what is at stake here is not how worthwhile it is, but whether it is love at all, which his view denies.

Second, I think the theory is psychologically problematic. Note that Kolodny is not saying only that an attachment is inappropriate when there is no belief that there is a historical relationship, but that there *cannot* be one, for an attachment is based on the belief that there is such a relationship – here is the quotation again: 'Special concern for a person is not love at all when there is no belief that a relationship renders it appropriate.' Indeed, he is bound to say that, since his theory states that, in good cases, the valuation of the relationship is what underlies the attachment and, in the case of absence of a belief in the existence of such a relationship, there cannot be a valuation of it. Therefore, if there is an emotional disposition that resembles an attachment, it must be underlain by a completely different standing mental state. But, of course, one might wonder, irrespective of the issue of appropriateness, whether an attachment in the presence of a long relationship and unrequited love (or 'unrequited friendship') are as distinct mental phenomena as this view renders them to be.

There is another sense in which the theory is psychologically problematic. It is silent precisely on what happens when an attachment is formed or when an attachment becomes stronger. All that Kolodny says is that there is a non-insistent reason to form an attachment, without elucidating what determines the agent to form the attachment based on that reason, or, in other words, it’s fine either way, we don’t really care what the difference is between the cases. For him, it’s a ‘take it or leave it’ case that requires no further thought. But presumably it is precisely what is going on when forming an attachment that interests us and that would elucidate the phenomenon of attachments.

The third problem is that I am not convinced that one can appreciate and value a common history with someone independently of what they currently feel for that person. Indeed, it seems that Kolodny gets the explanation the other way around and that common history with someone is important to us precisely because we have an attachment to the other person, the common history representing and symbolising the way we developed this attachment.

To see this, imagine the following scenario: two people X and Y meet up for a coffee, they talk but find that they do not really care about what the other thinks or what the other’s problems are, they do not really like each other very much and get bored soon. Each of them wonders why they spend time with the other, thinking it is should be the last time they have a coffee together, when all of a sudden they both have the same thought: ‘Shit! I forgot to take into account all our common history!’ And, given that they do indeed have a long common history in which they very much enjoyed each other’s company, both change their attitudes and their emotions, feeling and behaving as best friends[16]. Assuming that their relationship has so far been of the kind that Kolodny assumes is valuable, this is what, according to him, should happen. Yet this is: a) psychologically unrealistic; b) weird, to put it mildly.

Another way to make the point would be by appealing to amnesia. Kolodny himself uses the case of amnesia to present his theory as superior to the properties view. He argues that his theory can explain why amnesia extinguishes love, as one doesn’t have the belief about the relationship any more, while the properties theory cannot, for the properties are the same.

[16] Just to clarify, their recollection that they should take into account common history is not an emotional one, they do not remember each other in the past in good light, with a kind of nostalgia. Rather, they recollect a fact, that they have a common history, just as a fact that they forgot to take into account and the emotions are based on taking this fact into account.

But I think the scenario is even more damning towards his view. Indeed, even though X has forgotten about her relationship with Y, she can very well be *reminded* of it, that is, someone can tell her ‘Look, you had this-and-that sort of relationship with Y, that involved this-and-that’. I reckon this won’t make her jump up and down out of love for Y, as Kolodny’s theory would suggest is appropriate. What does this imply?

First, it implies that the psychological mechanism that Kolodny describes doesn’t really exist. Indeed, his theory implies that there is a mental state (or a mechanism more generally) that makes X respond to a belief about a relationship being of a certain (valuable) kind with a disposition of concern. And if such a mental state exists, at least in some people, this should lead them to react in the amnesia scenario above with an instant reappearance of the attachment.

Second, I take it that if X reacted with an instant emotional disposition to being reminded of her relationship with Y, we would think that there is something wrong about her. Indeed, the image that naturally comes to (my) mind is that X would be like a ‘loving machine’: you plug in facts about her past, she produces loving emotions.

The fourth problem comes nicely after that last image and it is the problem of office love. To remind ourselves, the worry about office love is that, even if X is willing to treat Y non-fungibly, there might be a sense in which it doesn’t matter to her that it is Y himself that is at the other end of the relationship and not someone else, say Z. This is because X just wanted to have a (non-fungible) relationship with someone, and Y happened to be the best candidate. But the mental states of X are very similar as in the case in which it were Z, for fundamentally she is motivated by a form of valuation that converts beliefs about her past into an emotional disposition. There is nothing specifically about Y in the motivation.

Another way to put the problem is that X could say something like this to Y: ‘If I had had the same relationship with Z, I would care about Z in the way I care about you.’ And, importantly, this sentence is not a counterfactual about her mental state, in the sense that if she had had the relationship with Z, she would have had different mental states that relate to Z rather than Y. (If this were the case, the sentence would say ‘I would have cared about Z' rather than 'I would care about Z'.) Rather, it is a counterfactual about what she would feel and do in light of facts about the world, that are external to her motivation. In the counterfactual scenario, X would have the same deep mental states.

4.3. TAKE-HOME MESSAGES

I think the main problem with the accounts that I have discussed, the properties view and the relationship view, is that they try to rationalise the attachment, i.e. to see it essentially as a rational response to facts that are external to the agent. The goal seems to be to make attachments intelligible to rational agents as such. According to such a framework, X as an agent is faced with a set of circumstances: the qualities of Y, her common history with Y, perhaps some set of preferences that she has and so on. Under these circumstances, the question would be why a rational agent would have an attachment to Y, in other words, which of the circumstances would justify such an attachment. Any other person who is rational could then project themselves into X's situation and see why (or why not) she has this or that attachment.

But I think that this is the wrong approach. The relation to Y would then be something thoroughly external to X's psychology, something very superficial, a matter of circumstances that she happens to find herself in. To make sense of attachments, I think we need to realise that they are something deeper and, as I will suggest, something that is not rationalisable.

5. ATTACHMENTS AS DRIVES

I will now put forward my theory, which is that an attachment is a form of drive, that is, a fundamental standing mental state that manifests itself in emotions, and that in the case of attachment is directed at a particular entity in the world.

5.1. WHAT DRIVES ARE

Let me start by detailing what I take a drive to be. The idea behind forming such a theory would be to capture 'what it is that we care about deep down'[17].

First, some preliminaries. A mental state can be occurrent or standing. An occurrent mental states subsists only as long as it is conscious. Perceptions are a very good example of occurrent mental states. Standing mental states, on the other hand, can continue to subsist even when they are not part of consciousness. Beliefs are naturally thought to fall into this category: occasionally, I have as part of my consciousness the belief that water is H_2O, but it is often supposed that I continue believing that water is H_2O even when asleep. Standing mental states might also be unconscious, in the sense that they are permanently not part of consciousness and cannot be brought into consciousness. It is clearly open to debate whether there exist unconscious mental states, or even standing mental states, but I won't go into that here.

Now, one state might be a *manifestation* of another state: willing a means can be a manifestation of willing an end; a belief about an individual dog can be a manifestation of a belief about dogs in general. A mental state gives rise to its manifestations and these manifestations cannot continue to exist if the original state disappears. Even though perceptions give rise to beliefs, these are not manifestations of perceptions, as they continue to exist when perceptions do not any more. A belief about an individual dog that is formed in light of a belief about dogs in general should disappear once the latter disappears. If it doesn't, it is no longer a manifestation of the more general belief and has a life of its own.

I will then define a *drive* as a mental state that manifests itself in emotions and behaviour, thus being related to motivation, *and* that is *not* a manifestation of another mental state. Drives would also be standing mental

[17] There is an influence of Nietzsche (2006 [1887]) in this undertaking, and he seems to have an idea of a 'drive' (*Trieb*, *Hang*) that might be similar to mine. It's hard to engage with him in a very analytical way, as he doesn't really give definitions – it's more his spirit that I'm following. See Stern (2015) though for a view that Nietzsche does not have a clear concept of drives.

states, reflecting the fact that a deep concern can remain constant over time, even when we do not have specific occurrent mental regarding that concern.

To see better what I mean by this, let's think how we would try to discover what drives an agent has. Faced with some manifestations, we can make wildly different conjectures about their fundamental states. One might think, in a Nietzschean vein, that all our mental states are manifestations of one fundamental drive, the will to power. But even leaving aside ambitious conjectures like this, there is always the question of what we really care about, deep down, and what is just a superficial concern. For instance, if someone plays the violin, what is it that they care about that leads to this? Seeking enjoyment? A concern for violin music? For music in general? For culture in general? An ideal of constant self-improvement? A form of resentment towards their parents, who dislike classical music? By asking such questions, what we are essentially asking is which drives underlie the emotions and behaviour of the agent. If they *seem* to have a concern for classical music, but this concern was formed in reaction to their parents disliking classical music, we might be suspicious that their concern is not really a drive in itself, that the underlying drive is something like a resentment of their parents. If the resentment disappears and the concern for classical music disappears with it, we would be even safer to conclude that the drive, that is, what ultimately motivated them, was their resentment of their parents, and the concern for classical music was just a manifestation of that resentment.

5.2. ATTACHMENTS AS DRIVES

I want to claim that an attachment is a drive. In other words, the fundamental standing motivational state that underlies the disposition we associate to the attachment is a drive that makes reference to the person or entity the attachment is to.

To see better what I am trying to do here, let's compare this view with the previous two theories, the properties view and the relationship view. According to the properties view, the drive that is in play in creating the emotions and behaviour of X towards Y is a kind of valuation of properties, a valuation that leads to emotions. According to the relationship view, it is a kind of valuation of relationships, or of a type of relationships. None of these drives are about Y, and indeed this is why the two views end up having the problems that they do, especially regarding fungibility and office love. According to both views, it seems to be the same drive that leads X to form relationships with all the people she forms relationships with. On the view that I am proposing, there is no deeper mental state than the attachment to the other person or entity, that is, the drive that leads X to form a relationship with Y is about Y. This means that there is a distinct drive for each attachment that X has.

How is the attachment born, on my view? It is, of course, born in the background of existing mental states – it is not born out of thin air. This is obvious, as thinking retrospectively, it often makes sense why some people became friends, given their previous concerns and interests. The attachments are also born in reaction to properties of the other person or entity, whether intrinsic properties, or extrinsic properties, such as the interactions had so far with them or the fact of being related to them. Presumably, certain kinds of attachments, such as those to our parents, are born when we are very young in a way that is not perfectly intelligible to us as adults. But, importantly, once the attachment is born, it has a life of its own and is not dependent on the other mental states that gave rise to it. We can now see why the imprinting metaphor that Nozick puts forward makes sense – the creation of the drive is the imprinting.

I want to say a few words about what I take to be the strength of an attachment. One might think that an attachment is strong insofar as it leads the agent to have many interactions with the person they are attached to. I think this is not correct. We should see the strength of the attachment in terms of how much motivational power it has, in terms of how much it leads the agent to have emotions in light of it, and to perform the actions that these emotions indicate. Yet, these actions need not always involve becoming closer or spending more time with the person one is attached to, and might actually lead the agent to *keep a distance.* Let's look at a quote from Hannah Arendt's obituary of W.H. Auden:

> I met Auden late in his life and mine—at an age when the easy, knowledgeable intimacy of friendships formed in one's youth can no longer be attained, because not enough life is left, or expected to be left, to share with another. Thus, we were very good friends but not intimate friends. (Arendt 1975)

We might interpret Arendt's remark as suggesting that because they met too late in life, they were not able to form strong attachments to each other. But I think something deeper might be going on here: realising that they couldn't built up the natural interaction that one builds earlier in life, they might have *chosen* not to try to become too intimate. If one has a strong attachment to another person and realises that the relationship cannot fully materialise, one might prevent it at all cost from becoming a sham and the motivation for this stems from the attachment itself. The more important the other person is in one's life, the more one might want to avoid the attachment being expressed in some unnatural, even fake intimacy. Hence a very strong attachment need not lead to too many activities done together and might actually lead one to abstain from performing many such activities.

We can now return to our initial observations regarding attachments made in the second section and see why they make sense. First, it is clear whence the non-fungibility comes. If the attachment that X has towards Y is a drive that is directed at Y, there is no sense in which X could replace Y with Z. She could perhaps form an attachment to Z that is stronger than the one to Y, but this doesn't count as a replacement, for the motivation has a different source.

We can recollect from section 2.2 that there was a tricky case, in which Y dies and X forms a similar relationship to Z some years after. Does this count as a form of replacement? Kraut suggested that it is a matter of convention, but now we can see clearly how we can answer the question without appealing to convention. If X had an attachment to Y and now has an attachment to Z, then it does not count as a replacement. (I'm not saying that there aren't other ethical worries about this, just that it's not a form of replacement.) If, on the other hand, X had some other drive that lead him to form both the relationship to Y and the one to Z, as is postulated by the properties and the relationship views, then it is a form of replacement[18].

We can now also give an account of what office love is and what might be problematic about it. In the case of office love, the drive that motivates the agent is a kind of commitment to a type of relationship (a commitment more or less à la Kolodny), and even if X treats Y as non-fungible, there is still a sense in which deep down, it doesn't matter whether it is Y or someone else, for what matters to X is the type of relationship. Paradoxical as it might sound, X wants to treat someone non-fungibly, but it doesn't matter whom she treats non-fungibly. According to my proposal, in genuine attachments, what X ultimately cares about is Y, the individual Y, not some type of relationship or some kind of office.

Finally, coming back to good old Bernard Williams and his wife, we can clearly see why acting from a drive as described above does not involve one thought too many. For the attachment that Williams has to his wife is individualised and is not a manifestation of or filtered through a general, impartial commitment that any agent might share. The fact that that he loves his wife is not, as it were, a consideration that weighs in his decision, a consideration that any other agent might understand and thus integrate into the reasoning process. Rather, the love for his wife is the *source* of the motivation for the action, the standpoint from which he reasons. Of course, the motivation to save his wife might compete with other motivations and might even be defeated (say in the case in which he can save either his wife or one million other people), but this competition is not a form of stepping back from his attachment and seeing what a rational agent should do, or something along those lines. Rather, it is a simple matter of which drive pushes harder, his attachment to his wife or some other drive.

5.3. THE EVOLUTION OF ATTACHMENTS

Let's recollect Amélie Rorty's point that people might expect from a relationship a complex interaction that changes the participants in

[18] It might be difficult to tell from a third person perspective what actually happens in such a scenario, whether it's the same drive that leads X to form relationships with Y and Z or not, but this is an epistemic problem, not a conceptual one.

meaningful way, and not merely mutual caring. We are now in a position to give an account of this.

I said that once an attachment is formed, it has a logic of its own, independently of the other mental states that lead to its formation, for even though other mental states contributed to its creation, the attachment is not a manifestation of any other mental state. Now, this 'logic' would include X's usual reactions to Y, her feeling happy when seeing him and anxious when Y expects some important news, her helping him when he needs it and many other, positive but also negative, emotions and reactions. But it might also include responses to new facts, to changes in the properties of Y, or to new events, which make the attachment evolve. Let's see how this works.

The key idea is that an attachment contains in itself its own potential for evolution. After the attachment is created, X finds out new things about Y, yet her reaction to these findings is not similar to the reactions she would have were she to observe the same characteristics in a random person. Rather, they are a manifestation of the attachment. Having bonded with Y by discussing their love of nature, she might then discover that Y has a penchant for garish watches, a penchant that she wouldn't have guessed. While the same discovery in a random person might have left X indifferent, if not slightly dismissive, she might find Y's penchant a delightful flaw. She might even start teasing Y about this and this might become a ritual for them. This is a sign that X's discovery of Y's penchant and her reaction to it have become embodied in the attachment. The attachment has evolved and now incorporates in some way Y's watches, which now have a significance in the relationship[19].

Just as new information about Y can be incorporated in X's attachment to him, so can common history. For instance, X and Y might have played a lot of chess together. Even if X is no longer interested in chess, it still has a special meaning for her. Seeing a chess set, she might suggest that they play 'just like in the old times', something she wouldn't suggest to anyone else. We can now see that common history is not, as

[19] Coking and Kennett have a similar example of Judy teasing her friend John for always wanting to be right (1998, p. 505).

Kolodny argues, the basis of the attachment; instead, it gains significance within the attachment and is thus embodied in it.

The attachment might also create other drives. In the second section, I said that X might have become interested in the violin in virtue of her attachment to Y, an interest that could survive the loss of attachment. If that is the case, that means that the attachment created another drive that encapsulates X's interest in the violin, that this interest is not (any more) a manifestation of her attachment to Y.

I think it is all these complex changes that Rorty is looking for. If an attachment is prone to such evolutions, I shall call it a *dynamic* attachment. We can contrast such a dynamic attachment with the attachment that parents might have towards their children. Many of these attachments are less prone to change and for this reason we might call them *static*. Importantly, whether an attachment is static or dynamic is not so much a decision that the agent makes irrespective of the attachments themselves (for instance, by waking up one morning with the thought that 'from today, I will make all my attachments dynamic'). Rather, given that the changes are driven by the attachment itself, it is a fundamental characteristic of the attachment whether it is more static or dynamic.

5.4. THE VALENCE OF ATTACHMENTS

I said at the beginning that I want to give an account of attachments, irrespective of whether they are positively valenced (love) or negatively valenced (hate). We can now see that the theory I am proposing fits well with this goal. Indeed, a drive that is directed at one person, say Y, can lead to positive emotions and helping behaviour or to negative emotions, such as resentment. Moreover, we can also have cases (I suppose the majority) in which there is a mixture of positive and negative emotions.

Importantly, I am talking about the valence, not the value of an attachment. It might be claimed (though I don't think that's true) that the more positively valenced an attachment is, the more value it has. Frankfurt would, I guess, hold something like this (2004). Someone like Rorty, on the other hand, might claim that it is better to be annoyed at your friend's clumsiness and irritated by their shrill voice rather than delight dumbly in all their traits, irrespective of how they are. I will discuss the evaluation of attachments in chapter 6.

I now want to show how the valence can change in an attachment. Let’s take the following case: suppose that X and Y bonded over their passion for 19th century industrial architecture; Y might then say to X the following: ‘I like the idea of unconditional love, but frankly I wouldn’t like you to love me if I were to become one of those people who demolish these beauties. I just don’t like the idea of you possibly loving such a person.’ What Y hopes is that X’s attachment to him cannot incorporate that potential dramatic change. But now suppose that X actually has such an attachment to Y as Y hopes she has, an attachment that cannot easily incorporate the dramatic changes. What happens if the dramatic change occurs and Y becomes what he dreaded to become? Does the attachment X has to Y just disappear? I think that a better explanation is that actually the loving attachment transforms into hate. X ends up hating Y – individually, that is, not as a member of a group. This transformation is the result of the significance Victorian industrial architecture had in the attachment before Y changed. The immense significance they attached to something they both strongly believed in made them vulnerable to hating each other. Hence, in many cases in which we are tempted to say that love has been replaced by hate, we should rather say that love *has metamorphosed* into hate – it is *the same drive*. In forming such an attachment, one binds one’s life to another person in a way that is deeper than any valence. I think this is a nice cherry on the theoretical cake that I’ve been baking.

The upshot of this discussion is that, unlike in static attachments, in very dynamic ones an aspect of that relation can be of utmost importance, and the loss of that aspect can transform love into hate. This is why parents rarely end up hating their children, irrespective of how dreadful they become, while the same cannot be said for strong relationships between adults[20]. So what can we say about the merits of such extremely dynamic

[20] The fact that parents tend to continue loving their children despite dramatic changes might be explained as the result of evolution as well as development of culture (Spartans were slightly less accepting of their children’s faults than contemporary Europeans are). I am not providing a competing explanation that undermines these. What I am claiming is that these theories would explain the fact that parents have

attachments of love that can transform into hate? On the one hand, there is something very deep in such attachments, a sense of importance of how things actually are that is not possible in more static attachments. On the other hand, there is always the danger that what is held so precious that its destruction can change love into hate might be a very superficial trait of the other, or indeed even an imagined one.

5.5. OBJECTIONS

I will now address some potential objections to my view.

First, one could claim that on my view the person having an attachment is not acting rationally when acting motivated by that attachment. Indeed, according to my view, if someone is about to save his wife as opposed to a stranger, his reason for doing this would be something like 'Mary is in danger'. Full stop. The attachment that he has to Mary is what leads him to take it as a reason, but the attachment itself does not figure in the reasoning process. Neither does their relationship. Yet people find this problematic. Here is Kolodny:

> The question, then, is whether that she is Mary is a normative reason for the agent to help her over a stranger who is at least as needy. It should be clear that it is not. The thought that she is Mary simply identifies a particular with itself; it does not ascribe a property to that particular that might make a certain response to it appropriate. (Kolodny 2003, p. 159)

The problem with this worry is the assumption that all things agents do should be based on normative reasons, which is exactly what I am denying. Indeed, my account implies that people who have attachments do not always act on normative reasons. Kolodny relates his worry to justification of one's gestures towards third parties:

relatively static attachments to their children.

> When called upon by a third party to justify his actions, Mary's husband will know better than to say, "She is Mary." He will know that he needs to convey that she is his wife. (Kolodny 2003, p. 159)

Well, he might say that it is his wife, but I take it that this does not so much provide the normative reason as convey his motivation, based on the accepted assumption that a man loves his wife. He could just as well have said that he loves her, with his love being not something to be taken into account when making the decision, but the source of decision itself.

As a second objection, my account might give the impression that if a parent does not have an attachment to her child, it is fine not to do anything about her child and just abandon him. Or, to take a more mundane case, if someone loses their attachment to her friend, it's all fine to behave as if they have never met. This is because, once the attachment is gone, the agent cannot have the same kind of motivation as before.

In response to this, I want to note that, besides the attachments, there might be duties that flow from various interactions or from the fact of being a parent. A person might be motivated to act from duties irrespective of their having an attachment to the person the duties are owed to. So in the cases above it's not all right to ignore these relationships, but still, anything one does is *not* a manifestation of an attachment, for there is no such attachment. The motivation would be essentially impersonal, of the kind that the hardcore deontologist has when visiting his friend in the hospital. All this might be sad, but it is as it is. If there's no love, there are no actions done from love.

The third objection is a related one. There is a kind of expectation of constancy in attachments. If X has a strong attachment to Y, Y might expect that unless something bad happens, X will continue having her attachment, that she won't come one day saying 'oh, my attachment has just vanished this morning, so I'm afraid this is the end'. This expectation would involve not so much the fact that the other party behaves in a dutiful manner, but that the very attachment continues. And clearly Kolodny is better situated than me to account for such an expectation. For, on my view, if the attachment is gone, it's gone, there's no way to reason the agent back into it.

I think our expectation has to do with the fact that we see the human mind as having a certain constancy, that the way it evolves is in some way dependent on the mental states that were there before. Drives do not just pop in and out of existence all the time. Without an expectation of this kind, we would struggle to make sense of individualised human relationships. So if someone seems to lose an attachment, we might think that there was something wrong with him. Perhaps there was no attachment in the first place. Or perhaps there were other drives that pushed the attachment out of existence. But I don't think that the expectation to keep the attachment is a normative one or, in other words, that we expect the agent to realise that this is the right thing to do and just do it. It's just something that usually happens.

A fourth and last objection might question the existence of the individualised drives, as I have framed them. 'This all sounds very nice, my dear child, but alas, these are just fairies.' The drives would be impersonal, and whatever relationship we had, it would just be a manifestation of an impersonal drive.

It's of course very hard to respond to this. It is to some extent an empirical question. But I think that if these individualised drives do not exist, what seemed most important in individualised relationships would turn out to be an illusion and we should feel extremely let down.

6. CONCLUSION

I hope to have shown that it is wrong to construe attachments as in some way rationalisable responses to circumstances that the agent faces: they are neither, as the properties view claims, responses to valued properties, nor, as the relationship view claims, responses to valued relationships. Instead, attachments are drives, that is, mental states that encapsulate what the agent cares about deep down and that manifests themselves in emotions. The drive that an attachment consists in is directed at the person (or entity) the attachment is to, meaning that the attachment is not underlain by a deeper concern of the agent than the person she is attached to. Besides solving the problems about attachments being *de re* and about acting directly from an attachment, the view proposed allows us to see how an attachment can be more or less dynamic.

CHAPTER 2. EMOTIONS

1. INTRODUCTION

In the previous chapter, I have given an account of what attachments are, arguing that they are drives, that is, categorical standing mental states that manifest themselves in emotions and behaviour. The drives that constitute our attachments are directed at particular entities, the entities we are attached to. Given that our attachments manifest themselves in emotions, the next thing we need to do in order to develop our understanding of attachments is to put forward a theory of emotions, explaining what emotions are, what their role is and how attachments manifest themselves in emotions. In this chapter, I will defend an account of emotions, essentially a variation on Julien Deonna and Fabrice Teroni's attitudinalist theory of emotions, which claims that emotions are bodily attitudes directed at a content (Deonna and Teroni 2012, 2015).

In the second section, I will provide some desiderata that a theory of emotions should account for. Yet I will differ from the majority of writers on the subject (D'Arms and Jacobson 2000; Deonna and Teroni 2012; Tappolet 2016; Müller 2017) in positing that at least the emotions that are manifestations of attachments do not have fittingness conditions, something that I will argue for in the third section. In the fourth section, I will examine what I take to be the most promising three accounts of emotions on the market and argue for the attitudinalist theory of Deonna and Teroni, while in the fifth section, I will explain the connection between emotion and action in light of the attitudinalist theory. In the last section, I will round off the discussion by connecting what I expound in this chapter with attachments and arguing that emotions make actions intelligible for the agent performing them and are therefore essential for self-understanding.

2. DESIDERATA FOR A THEORY OF EMOTIONS

The following desiderata are intuitive and, I take it, to a large extent accepted by people writing on emotions. It seems that any theory should either account for them or explain them away.

The first one is that emotions are intentional or, in other words, that they are about something. Fear can be directed at a dog or at the impending crash of the stock market. Of course, there's bound to be controversy about whether emotions are directed at an object or at a 'content' and of what kind this content might be (whether it has to be propositional, conceptual etc.), but this will not be, I hope, all too relevant for what I will say. I will use 'object' and 'content' more or less interchangeably. Also, the objects of emotions need not exist: one can have emotions towards entities one believes exist, but which do not actually exist (e.g. the bogeyman) or towards entities one knows do not exist (e.g. fictional characters)[21].

Second, there is a sense in which the intentionality of emotions is *borrowed* from another mental state (Deonna and Teroni 2012, p. 5). In less metaphorical terms, the objects of our emotions come into our mind via other mental states: in order to feel contempt for a man, one needs either to have some belief about him, or some thought or imaginative episode, or to sense-perceive him (or to have another mental state directed at him.) One does not feel contempt for a man out of thin air, without having any other mental state about him. We can thus contrast emotions with other mental states for which we do not have this requirement: in order to perceive something, we don't need to have a previous mental state about that object; perhaps the same holds for imagination. The mental state which brings the object of emotion into mind is called the *cognitive basis* of the emotion.

As a third desideratum, emotions have a distinct phenomenology that a theory should account for. I will divide the phenomenology into two

[21] The latter is more contentious – even though it seems intuitive that we have emotions towards fictional characters, some have doubted that these are genuine emotions (e.g. Walton 1990). However, these theories were based on the idea that emotions are linked to beliefs. Nowadays, it is common to assume that we do have emotions towards fiction – see Friend (2022) for a survey.

aspects, the first one related to bodily feelings and the second one to the fact that emotions 'colour the world'.

The first aspect is that emotions seem to be accompanied by bodily feelings. Fear, for instance, seems to involve, amongst other physiological changes, quick heart-beats, alertness of limbs etc. In his classic article, William James (1884) claimed that emotions just *are* forms of awareness of these bodily feelings. But we can just accept the more moderate claim that they involve bodily feelings.

Peter Goldie objected that there are some emotions, such as pride, that do not involve bodily feelings (2000, p. 52). While we can easily associate fear with a racing heart, there does not seem to be an equivalent physiological change in pride. A similar worry might be formulated regarding admiration – what are the physiological changes specific to admiration? Actually, positive emotions seem more difficult to associate with physiological changes than negative ones (Shiota et al. 2011).

In response to this, we should not assume that just because an agent cannot easily locate the bodily changes in an emotion, this means that they do not exist. We become more easily conscious of bodily changes when they are precisely localised and intense. If they are less intense and more diffuse (i.e. spread throughout the body), it might be more difficult to tell that we *do* experience bodily changes, let alone describe them. This does not mean that they are not part of consciousness, just that we do not report well conscious phenomena that are not very intense. If we concentrate on what happens in pride, we might identify an increase in energy or a 'warm chest' and this is also what subjects of admiration actually reported to have felt in an empirical study (Algoe and Haidt 2009). There is also empirical evidence of physiological changes involved in pride: Fourie et al. have found, by measuring the cardiovascular activity of subjects, 'a somatic SNS [sympathetic nervous system] arousal pattern for pride' (Fourie et al. 2011, p. 893), but this response might not correspond to the kind of localised feeling that we can easily translate into common parlance.

Moreover, we shouldn't construe the idea of bodily changes in too narrow a fashion, as a change to a particular part of the body. A change of posture, for instance, also counts as a physiological change. The change of posture is felt, it is not merely a matter of our body being distributed in space in a new way; our muscles tense in a different way, we experience the proprioception of our body differently. And it seems that something like this happens, for instance, in pride: unlike in shame, in which our shoulders droop, in pride we straighten up, we feel our chest pulling forward and a certain warm glow around our body, as if we are radiating[22].

The idea of physiological changes associated with emotions is not so much that there is a very precise bodily reaction involved in emotions, but that emotions are not just in the brain. The rest of the body, from the neck down, as it were, is also involved. It's a fairly minimal claim, but one that should be accounted for by a theory of emotion.

The second aspect of the phenomenology of emotions is, in metaphorical terms, that they 'colour the world'. What this means is that the object of emotion is not experienced neutrally, with the emotion on the side (Goldie 2000, pp. 59-60). If one is afraid of a dog, one does not see the dog as if one were not afraid and have fear on top of that. Rather, the dog appears in light of one's fear: its teeth and claws dominate its general image, which is one of fearsomeness. Again, this is somewhat vague, but at this point this is not a bad thing, as any attempt to spell out further what happens amounts to starting to build a theory.

The fourth desideratum relates to the connection between emotions and action. Roughly speaking, there seems to be some relation between emotion and action, a relation that is also visible in animals. Emotions are often used to explain action: when asking why someone ran away from a scene, 'she was afraid' usually amounts to a satisfying explanation.

The first thing to say in this respect is that there are some standard actions associated with emotions. Fear, for instance, is usually associated with flight from the object of fear. Yet sometimes fear leads to different reactions, such as fighting or freezing. So, there is not one definitive action for each emotion, but still, there is a set of actions that are specific to that

[22] This might have to do with a possible connection between the emotions we feel at this moment or the emotions we tend to feel and our 'demeanour', understood as 'expressive characteristics of bodily movement and posture, facial expression and voice *… as integrated and unified into an overall expressive manifestation of a whole person's being*' (Steward unpublished manuscript, her emphasis).

emotion and that seem normal in light of that emotion and some that do not seem so.

Again, we have some emotions that do not have clear actions associated with them. Pride and admiration are usual suspects. But still, there are actions, understood in a broad sense, associated with these emotions. While in shame, one tends to avoid the gaze of others, in pride one is prone to search the gaze of others, to expose the object of pride. In admiration, one is prone to spend time looking at the object of admiration.

All these are very basic, unsophisticated, almost automatic reactions that we associate with certain emotions. But this is not the whole story. There are many complex reactions that are intelligible in light of emotions. For instance, a proud person might do what is known as 'humblebragging', pointing to one's imperfections with the goal of appearing better overall. This action is related to pride and can be explained by pride; therefore, a theory of emotions should also account for how an emotion can lead to all sort of complex and culturally specific actions like this one.

The fifth desideratum regarding emotions is about 'their triggering [being] to a large extent automatic or involuntary' (Deonna 2006, p. 29). The word 'involuntary' here can be misleading: it is true that there is no sense in which an act of will precedes an emotion. We don't decide to feel sad and then feel sad. Yet we shouldn't assume from the outset that this means that in some sense 'they happen to us', just as pain happens to us when we are injured. As I have argued in the previous chapter, they are, at least sometimes, manifestations of the drives that represent what we care about. Regarding their being 'automatic', the idea is that we sometimes have an emotion before we process what is happening in a situation and pondering what this might mean for us. If we see a bear, we instantly feel fear, before we think that that is a bear and before we reason that there is a danger that the bear represents.

I take it that most writers on the topic (De Sousa 1987; D'Arms and Jacobson 2000; Deonna and Teroni 2012; Tappolet 2016; Müller 2017) think there is another desideratum, namely that emotions can be more or less fitting, or appropriate. For example, an instance of fear would be fitting insofar as its object is dangerous (or fearsome), where dangerousness (or fearsomeness) are assumed to be evaluative properties. I don't take this as a desideratum, because I think it is false that all emotions have fittingness conditions, something I shall argue for at length in the following section.

3. EMOTIONS WITHOUT FITTINGNESS

The concept of fittingness is usually introduced via examples: fear of a puppy is taken to be unfitting because puppies are not dangerous (or fearsome). The thought is that emotions aim in some structural way to respond to some value, to latch on to that value and hence that they are fitting to the extent that their object has that value. A typical way to cash this out is by talking about the 'formal object' of the emotion, which in the case of emotions is the value property that is common to all instances of an emotion of a certain kind – for instance, in the case of fear, the formal object would be 'dangerousness' (or 'fearsomeness'), in the case of admiration, 'admirability', etc.[23]

Following standard practice, I take fittingness to be an internal norm of emotions, essentially their 'correctness', and hence distinguish it from other types of goodness of emotions (D'Arms and Jacobson 2000). Accordingly, it might be immoral to experience a certain emotion, such as amusement at a funeral, even though it would be fitting in that there is something amusing. Similarly, it might be prudent not to be afraid of an imminent danger because fear might paralyse you, but that doesn't mean that fear would be unfitting. So unfittingness would not be simple badness, but rather a kind of 'mistake'.

In this section, I will argue that at least a large class of emotions do not have fittingness conditions and hence that it is not in the nature of emotions that they have fittingness conditions. This will serve as a basis for adjudicating between theories of emotions in the next section.

Now, as far as I know, there hasn't been much written against the idea of the fittingness of emotions. There has been a half-hearted argument against this thesis (Dokic and Lemaire 2013, 2015) based mostly on a

[23] Though (as per Teroni 2007), the formal object might not individuate emotion types – contempt and shame might have the same formal object.

purported failure of theories of emotions that aim to account for their fittingness. Another argument against the notion of fittingness (Lemaire 2014) is based on the idea that, unlike in belief, a purported norm of correctness is not transparent in forming emotions.

My strategy will be rather different: I will look at the emotions we have in the context of attachments, that is, of personal relations to people (or things more generally) about whom we care as individuals, not as members of a class (human beings, business partners, etc.). As argued in the first chapter, an attachment manifests itself in emotions that we have towards the other person, emotions that we wouldn't have towards a random person: we are happy when we see them, hopeful for some of their plans as if they were ours, melancholy when life distances us geographically, and finally we grieve when they pass away. I will argue that these emotions do not have fittingness conditions[24].

Given that we find it natural to have such emotions towards people that are close to us and not towards strangers, it is common to think that the fittingness conditions of these emotions have to take into account the relation that we have to the other person. For instance, grief is thought to be appropriate if the death is *a loss to me*, where this is often cashed out in terms of how important the relation to the deceased was (e.g. Nussbaum 2001, Cholbi 2017). Other emotions can be given a similar analysis.

We thus have the following tentative analysis: the fittingness of emotions involved in attachments depends on two factors: one is the *situational factor* – i.e. what has happened, in the case of grief 'that the person has passed away' – and the other is the *personal factor* – i.e. a feature of the relation that captures how important that relation is, a feature to be determined.

Following Teroni (2016), we can account for how the personal factor fits into the structure of an emotion in three ways. We start from the distinction between the 'actual object' and the 'formal object' of an emotion: the former is the thing in the world that the emotion is about (a dog, the stock market etc.), while the latter is the evaluative property that the actual

[24] Again, this does not mean that we cannot criticise this emotions – for instance, because they reflect some immoral attitude – just that they do not have this internal norm of 'correctness'.

object is supposed to instantiate if the emotion is fitting (e.g. dangerousness or fearsomeness in the case of fear). Now, the first way to account for the existence of the personal factor is by arguing that the actual object of these emotions involves the relation to ourselves, that it is in some way indexed to the emoter, while the evaluative concept that the emotion is supposed to latch on to would not be indexed to the emoter. Thus, if in grief, the evaluative property is 'being a loss', and even though we say that we grieve John, actually, what we would grieve would be 'John as my friend', to which the property of 'being a loss' would apply. The second way would be to argue that the formal object or the evaluative property is in some way indexed to the emoter, while the actual object is not. In grief, the formal object would thus be 'a loss to me' rather than 'a loss' *simpliciter*, while the actual object would be 'John'. As a third option (though this is more controversial), we can claim that even if there is no reference to the self either in the actual object or in the formal object of the emotion, the fittingness conditions can involve the self. This would not be unique to emotions and would apply to the fittingness of other mental states – the belief that 'It is raining' would be true if and only if it is raining *here*, even if there is no sense in which the current location is part of the content of the belief or the mode (which is simply 'belief'). Despite the differences between the three options, the fact remains that in all of them the fittingness of the emotions is dependent on the strength of the relation we have to the other person.

Now, the question is, what does the personal factor consist in? In the first chapter, I have argued that an attachment consists in a drive that manifests itself in emotions and behaviour. Yet one could still claim that the emotions manifested by the attachment should fit the historical interaction with the other person. I will examine this option first and then proceed to the more promising option, that the emotions should fit the attachment itself, that is, the drive itself. I will argue that both options lead to contradictions and hence that the idea of fittingness of these emotions should be dropped.

3.1. THE PERSONAL FACTOR IS THE HISTORY OF THE RELATIONSHIP

According to the first option, the personal factor that the emotions aim to fit is the history of emotions towards and interactions with the other person[25]. If I no longer have any occasion to see someone with whom I have had a good but not very close relationship, it would be fitting, the thought goes, to feel some sadness, but surely not despair. Despair would be fitting if I were to realise that I would never see my best friend again.

The problems with this view parallel the objections that I raised towards Kolodny's theory of attachments in the previous chapter.

The first problem with this view is that it focuses too much on the past and not enough on the present. When reacting to an acquaintance's being away for long, what intuitively matters primarily is whether we care about them *now*, not whether we cared about them one year ago. Even in grief, which is paradigmatically a past-directed emotion, there is a sense in which the present relation matters more. Indeed, supposed that one is pressed by a critic to say exactly what it is that makes one grieve after a person to whom one has had a minimal relation. One is unlikely to take a step back and concede the unfittingness and instead would say: 'Dunno… I kinda just miss her… I now realise that I care about her very much.' It doesn't matter whether this implies that one has just started to care about the deceased, after she passed away, or that one started to care earlier on and has just *realised* that. What matters is that one cares about the deceased *now* and that is why one grieves.

The second problem relates to the start of the relation. Being the start of the relationship, by definition one hasn't had any friendly emotions towards the other person, so there seems to be no personal factor a first emotion might fit. The fittingness view would then have the strange implication that relations cannot quite kick off.

The third problem is a generalisation of the second one. The worry is that according to the fittingness analyses, relations should essentially remain

[25] I use the phrase 'emotions aim to fit' non-committally, just to refer to what the fittingness conditions are dependent on.

the same throughout time. This is because any new emotion should aim to fit the personal factor, which consists in the previous emotions and interactions, so should be in the same ballpark as the previous emotions. But this is obviously absurd, as relations develop, becoming stronger or weaker, or change in various ways. In feeling more intense sadness than we would have expected at an acquaintance's being away or more joy at their success, this is not so much an instance of unfittingness, but rather a sign of what is very nicely called 'growing closer'. Similarly, feeling less of these emotions would be a sign of 'growing apart'.

In reply, the fittingness theorist could attempt to account for the fittingness of emotions that are stronger than those experienced before by appealing to the duration of a relation. The personal factor would involve not only the quality of the history of the relationship, but also its length. For instance, once we have felt moderate joy for a long time at an acquaintance successes, it would now be fitting to experience more joy. But this reply is a non-starter. It would have the most peculiar implication that if we are on good terms but not really friends with someone, yet we see them every day and feel emotions for them, it would be fitting to feel stronger and stronger emotions as time passes. But clearly, there seems to be no problem in our relation remaining the same throughout time.

Last but not least, once we have the theory of attachments as drives presented in the last chapter, we can see that this account of fittingness has further problems. Indeed, the account implies that what is fitting for an agent to feel in the context of one of her attachments is fully determined by the history of the interaction with the other person and the current situation. It follows that the attachment plays a largely insignificant role: the emotions it should manifest itself in are already determined by the history. It seems plausible that someone who does not have any attachment to the other person and who merely cares about human relationships in general could very easily have the required emotions.

The take-home lesson is that there is a certain *spontaneity* in the relations developing, a spontaneity that just cannot be captured by this version of the fittingness theories. The moments in which we grow closer to or grow apart from a person involve having emotions that do not fit in the sequence of emotions that we've had so far. This is the moment in which the drive that the attachment consists in changes, becoming stronger or weaker.

3.2. THE PERSONAL FACTOR IS THE ATTACHMENT ITSELF

Taking our lead from the failure of the previous attempt, we could try to cash out the fittingness of emotions in terms of what relation we have to the other person *at this moment*. In other words, the personal factor that determines, together with the situational factor, the fittingness conditions, is the attachment, that is, the drive itself. To put some flesh on the bones, the fittingness theorist would presumably need to say a bit more about what makes an emotion fit a certain drive. But irrespective of that, I will provide three arguments suggesting that this approach is not really promising. None of the arguments is individually devastating for the fittingness view, but together I hope they would constitute a strong case against it.

First of all, I have argued in the previous chapter that these emotions are manifestations of the attachment. The emotions encapsulate what importance the attachment leads the agent to assign to a certain situation. So why should we think that, in addition to being manifestations of an attachment, the emotions also aim to fit that attachment? This is an extra theoretical claim that does not seem in any way necessary. The motivation behind the idea that emotions have fittingness conditions tends to be based on the intuition that an agent can misreact to a situation – just think of the examples of fear, which is purportedly fitting only if its object is fearsome or dangerous. But the attachment is not quite part of the situation that the agent reacts to; rather, it is what drives the agent to react to the situation. The view that emotions aim to fit the attachment would work much better in a framework in which the emotions would *not* be a manifestation of this attachment. In that case, the attachment would be something 'exterior' to the emotions, part of the situation the agent is responding to, and one could therefore naturally suggest that the emotions aim to fit it just as they might aim to fit the rest of the situation. But given that the emotions are manifestations of the attachment, it seems at best unnecessary to posit that they aim to fit the attachment.

Of course, the emotions would *de facto* be fairly representative of the drive they are caused by. If I am friends with someone, my emotions towards them would presumably reflect that fact because they are a manifestation of the attachment to that person. Yet given that the attachment is the source of the emotions and not part of the situation the emotions are a reaction to, we should rather say that these emotions *reveal* or *reflect* that attachment, rather than that they aim to fit it.

The second problem is related to the fact, mentioned in the previous chapter, that an emotion can be a manifestation of more than one drive or indeed of more than one attachment. Presumably the fittingness theorist would say that the personal factor would consist in the set of all relevant drives, attachments included. For instance, suppose that my friends Jane and John, whom I care about equally, play chess, and I am invested in their game. No matter what the result of the game is, I wouldn't have an extreme reaction, either of sadness or of happiness. Indeed, suppose that Jane wins. My attachment to Jane would impel me to be happy about the result, while my attachment to John would impel me to be sad, these two forces giving rise to some kind of ambivalent reaction. The fittingness theorist would probably be happy with this and would claim that the emotion is fitting in that it tries to capture the significance of the situation with respect to all the drives.

Now, here is the problem: even if the resultant emotion would in some way seem fitting, neither of the drives aims to give rise to an emotion that fits all the drives. Rather, each of them pushes, so to speak, in its own direction. So, if none of the drives attempts to create an overall fit and the resulting emotion is just the combination of various 'pushes', in what sense is this emotion aiming to fit all my drives? It seems that there is no mental process that actively attempts to provide the overall fit. And if there is no mental process that aims structurally to create emotions that are in some sense fitting, fittingness would be something that *just happens*. It would therefore be unnecessary to posit it as an internal norm of emotions.

Let's move on to the third and last problem. Some emotions do not only manifest some drive, but also signal some change in the drive. These would include presumably emotional reactions to new and relevant information about the other person. For instance, if I find out that my friend John has been lying to me about something, I will not only get angry at him, but this anger might signify a change in my attachment to him. This is an extreme example, being something that imperils my attachment to him, but similar things can be said about less dramatic discoveries, for instance that he likes badminton. Presumably, this might change the attachment somewhat, adding a new dimension to it, not necessarily in a good or bad way. Now, it would be natural, though perhaps not obligatory, for the fittingness theorist to claim that an emotion aims to fit something that is already there at the moment of its being triggered, or even a millisecond before. Yet, the attachment as changed in the emotion was not like this a millisecond before the emotion was triggered. So the emotion cannot aim to fit the attachment as changed.

To conclude this section, it's hard to argue that a mental state does *not* have a certain internal norm, but I hope to have shown that in the case of emotions that are manifestations of attachments, positing fittingness as an internal norm is unnecessary and also creates problems. Given that the case for fittingness is not so strong and usually based on the intuitions we have in cases that do not involve attachments, it seems sensible to conclude that emotions that are manifestations of attachments do not have fittingness conditions.

4. THEORIES OF EMOTIONS

We have reached the point when we should discuss various option for theories of emotions and decide which one we should prefer. I shall say from the outset that most of what I say in the thesis is compatible with the all the theories I discuss, yet I will still argue that one is better and use it in the rest of the thesis.

I will discuss three theories: the perceptualist theory, which claims that emotions are a kind of perception of value, the construalist theory, which claims that emotions are a kind of construal, and the attitudinalist theory, which claims that emotions are a bodily attitude directed at a content. I will argue that the third is the best theory of how to conceive of emotions in general and those that are manifestations of attachments in particular.

I won't discuss two classic theories of emotions: the Jamesian theory (James 1884), which claims that emotions are just instances of awareness of bodily feelings, and the judgmentalist theory (Solomon 1993 [1976], Nussbaum 2001), which claims that emotions are judgments of value. This is because I take it that the three theories that I am discussing are improvements on these, that capture their main insights while avoiding immediate problems: the perceptualist and the construalist on the judgmentalist theory, and the attitudinalist on the Jamesian theory.

4.1. THE PERCEPTUALIST THEORY

The perceptualist theory claims that emotions are a kind of perception of evaluative properties (Tappolet 2016). This view starts from the idea that emotions are in many ways similar to sense-perceptions, so that pursuing this analogy could help us elucidate the phenomenon of emotions.

The first similarity is that they are intentional, about objects in the world, and that in both mental states, the world appears in a certain way, with a rich phenomenology. In visual perception, objects appear in colour, while in the case of emotions, we have already used the metaphor that objects are 'coloured' by the emotion they are the object of. Of course, emotions can be about objects that perceptions cannot be about, such as world peace or mathematical theorems, but this shouldn't matter too much.

The second similarity is that, like sense-perceptions, emotions are very often automatic responses to the environment that are not mediated by judgments. Metaphorically, when an object enters our visual field, it strikes us with its presence; similarly, in emotions like fear, the dangerousness of an object can also be said to strike us. Even if we have to see the object or have some kind of mental image about it first, the emotion can precede any judgment that we make regarding it.

The third similarity, which perceptualists often lay some stress on, is that we have instances of both sense-perceptions and emotions that go in some way against our best judgments. In the case of sense-perceptions, we have the usual illusions, such as the Müller-Lyer illusion, in which we see one line longer than another, even though they are (and, indeed, we know them to be) of the same length. In the case of emotions, we have, for instance, phobias, in which we are afraid of something we know is not dangerous. The perceptualist can thus hope for a neat way of explaining phobias and other wayward emotions, by assimilating them to visual illusions.

So, emotions are *like* sense-perceptions. But they aren't sense-perceptions, thus the perceptualist needs a broader definition of perception that would render both sense-perceptions and emotions instances of it. One such attempt at a definition has been made by Christine Tappolet:

> According to a liberal, but plausible account, perception can be defined as a kind of awareness of things and qualities. Put metaphorically, perception is a form of openness to the world; when things go well, what we are aware of is a fragment of the world. As far as I can see, the features that are most important on such a liberal account are among those that emotions share with sensory experiences: phenomenal properties, automaticity, world-guidedness, correctness conditions, and informational encapsulation. (2016, pp. 29-30)

There are some immediate problems with the idea of emotions being a form of perception understood as 'a form of openness to the world'. First, as said before, we can have emotions towards non-existent entities, such as fictional characters, or more generally we can have emotions towards objects that are not present, such as the fear one might have of terrorists. This means that it is quite a stretch to say that we grasp something from the world in emotions. A second problem is that, again as mentioned before, there has to be another mental state, a cognitive basis, that serves as an intermediator between the world and the emotion (Deonna and Teroni 2012, p. 69). Even when we have emotions about something that is present, we need to have some kind of awareness of its presence, such as seeing it, in order to have an emotion about it. The emotion does not reach directly into the world. It follows that the property that would be ‘perceived’ in emotion, whatever that means, would essentially be ‘perceived’ in some sense within the content of another mental state. It could well be that this other mental state, if it is for instance a form of sense-perception, opens us to the world and allows us in some sense to be aware of a fragment of the world; in that case, an emotional reaction to what we are aware of in perception would also indirectly be a kind of awareness of the world. But the way Tappolet phrases her view, saying that the emoter is ‘aware of a fragment of the world’, obfuscates this fact, that either there is no awareness whatsoever, or the awareness is merely indirect, mediated by another mental state.

I think that we could come up with a better way to elucidate the perceptual analogy in the case of emotions by thinking of other similar phenomena: I look at a building and after some time, I recognise its grace; I visually remember how the dinner went last evening and suddenly it strikes me that the host was taller than everyone else; or I form a mental image of the heroine of a novel and then realise that she has dark hair. In all these cases, what happens is that in the content of a mental state, I ‘perceive’ a property (e.g. grace or height) of the object of that mental state. This form of ‘perception’, whatever it is, seems to be something over and above the state which provides the original content to which we react emotionally (cf. Dokic and Lemaire 2013, pp. 230-2). If we want to put forward a perceptualist theory of emotions, it seems to me that the best option would be to construe an emotion as a similar kind of ‘perception’ of a property within a content of another mental state. The kind of ‘openness’ that Tappolet talks about in the quotation above would be just a preparedness to respond in a certain way to contents presented in one’s mind by various mental states such as perceptions, beliefs and so on.

Framing the theory like this would allow the perceptualist to account for the fact that in emotion the world appears differently, where by ‘the world' I mean both what the agent sense-perceives and what the agent thinks about. Indeed, the 'perception' of the new property changes how the content is apprehended by the agent, just as in the other cases cited above.

Now, the question is, what kind of properties are 'perceived' in emotions? Tappolet wants to claim that in emotions the agent 'perceives' *evaluative properties* (cf. Döring 2007). This would also be what distinguishes emotions as mental states from the examples given above in which the agent 'perceives' some other kind of property in a given content: if I form a mental image of last night's dinner and perceive the host's being taller than the guests or if I regard a tree and perceive, within the content of my visual representation, its perfect symmetry, I am not having an emotion, for the properties of being tall or symmetric are not evaluative properties.

I have argued in the previous section that emotions do not necessarily have fittingness conditions, that is, that they do not necessarily attempt to latch on to a property. It follows that what is 'perceived' in emotions cannot always be some evaluative property that obtains in the world. The perceptualist could accommodate this by claiming that actually the properties 'perceived' in emotions are essentially projected, that we experience the world in terms of what we care about and this appears as a property of things[26]. However, I doubt that perceptualists would be happy with this move, as I take it that one key motivation is that emotions aim to latch on to some values 'out there'.

Moreover, it seems to me that this analysis of 'perception' does not render the emotions different enough from other mental states, as different, for instance, as sense-perceptions are from desires. Indeed, as shown above, there are other instances in which a property is perceived in a given content, such as when I perceive the tallness of the host in my mental image of last night's dinner. The perceptualist can differentiate emotions by saying that a special kind of (projected) property is perceived in them, but this doesn't seem to render emotions a distinct type of mental state; they would be just a special instance of a broader phenomenon. Of course, we cannot exclude completely this possibility, but we should prefer an account that renders

[26] As the wonderful Humean slogan has it, 'the mind has a great propensity to spread itself on external objects' (Hume 1960 [1739], p. 167).

emotions more clearly distinct from other mental states, capturing our intuition that they do form a distinct class of mental states, on a par with beliefs, desires, sense-perceptions, etc.

Another problem with the perceptualist theory is that it seems to render emotions a bit too detached. Intuitively, if one has an emotion towards an object, there is a certain involvement of the agent in the situation, the agent cares about the situation. But if we assume that emotions are just some kind of perception, it is not clear how we can really account for this involvement. Indeed, when we think of sense-perceptions, there is no involvement in them, they just happen to us.

This might be related to bodily feelings (cf. Deonna and Teroni 2012, pp. 68-9). It seems that the perceptualist has to say that bodily feelings are not really part of the emotions – how could bodily feelings be involved in a kind of perception? And we might connect the involvement that I have talked about above with bodily feelings – it is not for nothing that we say about people that they get het up about something. Of course, perceptualists can claim that emotions tend to cause bodily feelings, but this might be too thin a connection and would not account for the involvement of the agent in the emotion.

Overall, even if there is no damning objection, the perceptualist theory is not in a very good position. Once we accept that emotions do not necessarily have fittingness conditions, a central motivation for the theory disappears and the properties that are supposed to be perceived in emotions are rendered rather mysterious. Moreover, the perceptualists struggle to give an account of perception that works for emotions. If we go down the route I have proposed – i.e. claiming that emotions are a form of ‘perception’ of a property in a given content – that theory would render emotions too much like other similar kinds of ‘perception’. Lastly, perceptualists struggle to account for the bodily feelings associated with emotions in a satisfactory way.

4.2. THE CONSTRUALIST THEORY

I will now discuss a second promising theory, which appears to be less popular, partly because it is somewhat hard to pin down, yet which deserves attention. This is the theory of Robert C. Roberts (2003), a theory that I will label ‘construalist’.

Essentially, Roberts claims that emotions are a form of construal, namely a ‘concern-based construal’. To understand what this means, we need to take the terms in turn and explain what is a ‘construal’ and what is a ‘concern’.

Roberts says that in a construal, ‘one thing is perceived in terms of something else’ (2003, p. 76). His main attempt at explaining this concept is by way of examples (2003, pp. 69-75), which are supposed to get us to feel what he is talking about. There is the classic example of construing a certain drawing as young woman or as an old woman – here, we see a drawing in terms of a kind of woman. This is the same kind of example as seeing the duck-rabbit drawing as a duck or as a rabbit. Another example is that of construing the first signs on a page in a book as a header (and hence ignoring them). A third, very convincing, example, is taken from a famous passage in *Brideshead Revisted*. When Charles Ryder first sees Julia Flyte, he construes her as Sebastian’s sister and hence all her features are seen as either what marks her as part of the family or as what makes her unique. By seeing her as her brother’s sister, Charles Ryder experiences Julia in a special way (that perhaps anticipates what is going to happen in the novel).

When it comes to the concept of ‘concern’, Roberts says:

> I use ‘concern’ to denote desires and aversions, along with the attachments and interests from which many of our desires and aversions derive. Concerns can be biological (‘instinctive’) or learned, general or specific, ultimate or derivative, and dispositional or occurrent. (2003, p. 142)

The concept of ‘concern’ seems to come down to any form of caring about something. With the risk of being unfair to Roberts, I find this rather uninformative. The ‘things’ on his list do not seem to have much in common except that they are the kind of things that *could*, on some views, give rise to emotions. For instance, are they at least all mental states, or does he think they are? Anyway, we can safely assume that an attachment, as I defined it, is a concern. There would also be non-individualised concerns: perhaps one cares about justice, football or stamps.

Now, an emotion would be a ‘concern-based construal’ or, in other words, a form of experiencing an object in the world in terms of one’s concerns[27]. If I have a strong concern for my shoes and I see a dog intending to chew them, this would amount to a kind of danger and I might start to construe the situation in terms of the concern for my shoes and the danger they face. Various features of the situation would become salient: the distance between the dog and my shoes, the dog’s teeth as the symbol of the impending danger and so on. All this construal would just amount to what we call fear.

One positive feature of the theory is that it seems to capture aspect of the phenomenology quite well. As I said at the beginning, when we have an emotion, the world appears differently, even if what we sense-perceive remains the same. And this phenomenon seems similar to what happens when we change from seeing the duck-rabbit as a duck to seeing it as a rabbit – we see the same drawing, yet the phenomenology changes. Or similar to what happens when Charles Ryder sees Julia as Sebastian’s brother. So, it is natural to conclude that all these are the same kind of mental phenomenona that could be captured by the notion of ‘construal’.

However, I’m not sure the theory accounts for the other part of the phenomenology very well, namely the fact that emotions seem to involve bodily feelings. Of course, a construalist, just as a perceptualist, can say that emotions tend to be accompanied by bodily feelings, which would be why we associate bodily feelings with emotions, but, again, one might hope for a stronger connection.

Another worry that I have – and that I had regarding the perceptualist theory as well – is that the construalist view renders emotions

[27] Roberts’ theory could be thought of as a more sophisticated version of Bennett Helm’s, who argues that any emotion involves a target (the actual object) and a focus (which corresponds to the concern) (2010, pp. 57-8).

too similar to other mental states. As Roberts himself says, there are many kinds of construals and only a few of them are emotions, those that are concern-based. This goes against our intuition that emotions are a type of mental state that is clearly different from other mental states, just as sense-perceptions are distinct from desires. Perhaps there could be a way for the construalist to alleviate this worry, but I cannot see one for now.

A more significant worry is that there seems to be a way in which I can construe something in terms of my own concerns without having an emotion. To show this, I shall develop an example in a roundabout way. First, I can construe something in terms of a concern of *someone else*. Indeed, suppose that I have an acquaintance who is very interested in violin music. When I see (and hear) a good violin, say a Stradivarius, I might construe everything that I see and hear in terms of this: I might hear the sound in terms of what my acquaintance finds remarkable in violins, in terms of what she might think would sound best on this violin, etc. (Of course, I need to have significant *knowledge* to do this, but not a *concern*.) Now, in an alternative scenario, suppose that it is I who has the interest in violin music. In this scenario, it seems that I can have essentially the same mental states as in the first scenario, just that they are based on my own concern and not on my acquaintance's. What I am doing is essentially treating myself from a third-person perspective. Given the way I built the first scenario, I take it as obvious that the concern involved does not amount to an emotion, so, given that in the second scenario, the same kind of mental states are involved and there is only a difference of reference, there should not be an emotion in that scenario either. Yet Roberts' theory seems to imply that in the second scenario the construal amounts to an emotion. The problem with the theory is that the construal is too intellectual an activity and need not involve the kind of *involvement* specific of emotions.

Roberts could reply that in the scenario I described I cheated, because even though the construal is based on my own concerns, I am not really taking them into account as *my* concerns, that is, in treating myself from a third-person perspective, I essentially 'forget' that they are my concerns. The concern should enter the emotion in some way first-personally. I agree with this, but this would render his account uninformative, for then we are left with the following question: what is it to construe a situation in a way that reflects your concerns and not merely takes them into account? This question almost takes us back to the original question of what emotions are.

Another worry that I have regarding Roberts' approach is that, contrary to what he says, in emotions, the concerns that drive the emotions are not always part of the phenomenology of the emotion: if one is afraid of a bear, the set of concerns that give rise to the fear might include, of course, a concern to live, but can also include a concern for one's *magnum opus*, whose completion depends, amongst other things, on not being eaten by bears. It can also include a concern for one's children, who would suffer greatly in case their parent is killed by a bear. All these concerns can shape the fear and perhaps someone with no concerns whatsoever might not experience great fear in the face of dying. Yet, I take it as obvious that they are not present in one's mind when one is in the grip of fear.

Roberts actually claims that in general what the terms of the construal are need not be transparent to the agent (2003, p. 72). When one construes the first signs on a page as a header, one does so without realising. Or perhaps Charles Ryder could see Julia as Sebastian's sister without being aware of that (although he is aware in the novel): he is just too familiar with Sebastian's face such that the construal is almost automatic.

This reply might show that the person experiencing a form of construal might not be able to analyse and describe his experience. But this is weaker than my claim in the case of fear described above. In that case, my worry is that the source of fear, that is, the concern, might not make a phenomenological difference, which is a stronger claim than that whatever phenomenological difference there is, it is not noticed by the agent as such. When faced with a bear, whether it stems from one's concern for one's children or for one's *magnum opus* might not change the phenomenology of the emotion at all.

A further reply from Roberts, claiming that, in some cases, the concern might not even make a phenomenological difference, would not work. I take it that it is part of the concept of a construal that what the construal is based on makes a phenomenological difference. All of Roberts' examples of construals involve a phenomenological difference – seeing the duck-rabbit as a duck or as a rabbit changes the phenomenology and positing the notion of a construal serves to explain whence the change comes. If there is no phenomenological difference, I am not sure I understand the concept of a construal.

My last point against Roberts stems from my account of attachments in the previous chapter. Both he and I think that some emotions are related to our attachments (and other concerns) and it seems to me that one of the implicit motivations for his theory is that we need to account for this connection. But once we assume that these emotions are a manifestation of the attachments, there is no further need to account for the relation at the level of the mental state of the emotion, for we already have one relation, in one being a manifestation of the other. His theory would have worked better in the case in which attachments (or other concerns) were something in a way exterior to the agent (e.g. a history of interaction), and then there would have been a need to bring the object of emotion and the attachment together at the level of emotion. In that case, emotions would have fitted nicely in the list of examples of construals he gives, all of which involve bringing together two 'things' (objects, concepts, etc.) external to the agent, by seeing one in terms of the other, as for instance when we bring together a drawing and the concept of 'duck' in seeing the drawing as a duck (and not as a rabbit.)

Moreover, once we assume that emotions are manifestations of attachments, it is not clear why an attachment would lead to a mental state that involves a construal in terms of that attachment. The purpose of emotions is to reach out into the world, to interpret what is relevant for the agent, so there is no clear need for them to manifest their source as well. Of course, this is not strictly speaking an objection, but undermines some of the implicit motivation for the construalist view.

Perhaps a better notion of 'construal' and a more careful account of emotions as construals might solve some of the problems that I mentioned, but for now I cannot see how.

4.3. THE ATTITUDINALIST THEORY

We have reached the theory that I find the most promising and that I will argue for, the attitudinalist theory of Julien Deonna and Fabrice Teroni (2012, 2015).

Deonna and Teroni start from the assumption that we can analyse (at least some) mental states in terms of mode (or attitude) and content. In the case of the belief that it is raining, believing is the mode (or attitude) and 'that it is raining' is the content. A mode can be directed at many contents: one can believe 'that it is raining', 'that Paris is the capital of France', etc. Also, many modes can be directed at the same content: one can believe, suppose, imagine and, as we shall see, hope, fear, etc., 'that it is raining'. I will use 'content' quite loosely, to encompass anything that can be the object of a mental state: it can be a proposition, but it can also be an object (a dog, as when one imagines a dog), it might be non-conceptual, etc. Not much will hinge on what we take the content to be.

Now, in the perceptualist theory, what is specific to the emotion is located at the level of content. One 'perceives' a value, or a projected value, and this value is located at the level of the content of the 'perception'. In Roberts' theory, it is unclear if and how we should analyse the emotion in terms of mode and content, but I think it is more natural to assume that the object of emotion, the concern and the construal of one in terms of the other are located at the level of content.

Deonna and Teroni locate what is specific to the emotions at the level of mode (or attitude – hence the name 'attitudinalism'). In particular, they claim that to each type of emotion (fear, hope etc.) there corresponds a specific mode that is directed at various contents. Let's take the content and mode in turn.

First, the content. As mentioned before, Deonna and Teroni argue that emotions just borrow the content from a different mental state, which serves as a 'cognitive basis' for the emotion. It can be the content of a belief, for instance – as in the case of fear of the impending crash of the stock market, a fear based on the belief that this is likely to happen. Or it can be the content of a perception – one can be afraid of the dog that one sees even before forming any belief about the dog. Or it can also be the content of an episode of imagination – one imagines a remarkable scholar, for instance by reading a novel, and feels admiration for her. By following this route, Deonna and Teroni avoid making emotions too dependent on only one type of mental state (usually belief), a dependence that usually leads to paradoxes (for instance, the paradox of how we feel emotions for fictional characters.)

Second, and more interestingly, the mode. The claim Deonna and Teroni put forward, following the work of Edouard Claparède (1928), is, essentially, that the mode of the emotion is an attitude taken by the *body* towards a content. The bodily feelings that are associated with emotions should not be thought as distinct episodes, as in the 'curiously atomistic approach to bodily sensations implicit in many accounts of their role in emotions' (Deonna and Teroni 2012, p. 79). Instead,

> the emotionally relevant bodily changes are experienced as distinct stances we adopt towards specific objects. That is to say, we should conceive of emotions as distinctive types of bodily awareness, where the subject experiences her body holistically as taking an attitude towards a certain object (Deonna and Teroni 2012, p. 79)

So, the bodily changes are not experienced, as it were, in themselves, but as directed at a content. When we are afraid, it is true that our heart starts racing and our body becomes alert, but we need not be aware of these individual changes. Rather, our bodily reaction marks the object of fear as to be avoided. What we are very much aware of in such an episode of fear is the object of fear and the bodily reaction, as a whole, makes us be aware of this object in a special way, the way specific to fear. I take it that this is presumably the part that is hardest to swallow in their theory, namely, that bodily changes can be directed at a content.

Importantly, to understand the view, we should discuss Deonna and Teroni's answer to a worry of Peter Goldie's (2000, p. 59-60). As mentioned in the second section, Goldie claims, on the basis of his phenomenological analysis, that an emotion is not something added to a value-neutral content. A person who is not afraid of a dog and a person who is afraid do not experience the content of their, say, perception in the same way, with one having, on top of that, the emotion of fear. Rather, the fearful person experiences the dog in light of the fear. From these considerations, Goldie concludes that an emotion cannot be just an attitude directed at a value-neutral content.

The reply from Deonna and Teroni is essentially that, even though we analyse mental state in terms of attitude (or mode) and content, phenomenologically they blend, such that the agent does not experience the mode, on the one hand, and the content, on the other (2015, pp. 304-7). Rather, the phenomenology of the mental state is (in some cases at least) the result of both the attitude and the content. The content is not something that stays in consciousness, waiting for modes to be directed at it, like darts at a dartboard[28].

The main point in favour of the attitudinalist theory is that it captures the phenomenology very well. The two aspects of the phenomenology of emotions, namely the bodily feelings and the fact that emotions colour the world, are accounted for and rendered aspects of the same unified experience.

Moreover, we should also note that, even if we have very strong bodily feelings in one emotion, we often do not focus on them and only observe them when we pay attention to our body. For instance, when we fear an approaching bear, all our attention is directed at the bear, and we often only observe our racing heart after the danger is gone – 'I still haven't calmed down!' Or, even if we observe the bodily feelings during the emotion, the phenomenon seems to be a kind of 'stepping out' of the emotion, of observing it in a second-order mental state. We can usefully contrast this phenomenon with, for instance, bodily pain (or pleasure), which we often notice very easily and which directs our attention to the parts of the body that are in pain. Now, the attitudinalist theory fits perfectly with all this. Indeed, by claiming that the bodily feelings are directed at the object of emotion, it shows why our attention in emotions is directed at the

[28] My comparison, not Deonna and Teroni's. I don't know if they would be happy with it.

object and not at our bodily feelings. This discussion should hopefully alleviate initial worries about bodily feelings being directed at an object.

One objection to the attitudinalist view in the literature is that this view cannot account for the fittingness of emotions, as fittingness should be explained in terms of some content being correct (Rossi and Tappolet 2019). Deonna and Teroni have replied to this (Deonna and Teroni 2022), but given what I have argued in the previous section – namely, that at least many emotions do not have fittingness conditions and that it is not their purpose to look for value in the world – I think the objection has lost its force. Indeed, if there were some emotions that do have fittingness conditions, their having these conditions cannot be a consequence of their nature as emotions.

Even more, in order to make their theory account for the fittingness of emotions, Deonna and Teroni have contended that there is a correspondence, more or less, between emotion types and thick evaluative properties: shame with the shameful, fear with the fearsome, etc. (2012, pp. 40-2 and 80-5). In this way, there would be clearly separated types of emotions and to each type there would correspond one mode. Within one type, the variation would presumably be primarily in terms of intensity. Once we give up the idea of fittingness, there is no need to claim that an individual emotion has to belong to one of a handful of types (fear, anger, etc.), and we might indeed have a continuum of emotions, some of which are very hard to pin down. I don't want to commit myself to a view on this matter and will leave the various options open.

There is one last, important part of Deonna and Teroni's view, namely, the connection of emotions to action. I will dedicate the next section to this, as it is important for understanding emotions in general and for this thesis in particular.

5. EMOTIONS AND ACTION

Deonna and Teroni share the common intuition that there is a connection between emotions and actions, that what one feels has something to do with what one does (2012 pp. 78-85). Yet, it's quite a hard to pin down this idea.

Here, I find their position, even if on the right tracks, just a little bit vague and hope to clarify it.

Deonna and Teroni argue that the mode directed at a content involves (or even is) a feeling of 'action readiness'. The physiological changes involved in an emotion are those that prepare the agent for action. For instance, in fear, the increased heart-beats and general alertness of limbs prepares the agent to run or to defend herself from the dangerous object. Yet again, these changes form a global attitude directed at an object that is, typically at least, outside one's body and hence we experience the action readiness as directed at an object:

> In fear, the relevant action readiness should be described as follows: we feel the way our body is poised to act in a way that will contribute to the neutralization of what provokes the fear. In anger, we feel the way our body is prepared for active hostility to whatever causes the anger. In shame, we feel the way our body is poised to hide from the gaze of others that typically causes the shame. (Deonna and Teroni 2012, p. 80)

Following Nico Frijda, Deonna and Teroni claim that in an emotion, these bodily feelings are directed at the object of emotion, and this object starts appearing in light of these instances of 'action readiness':

> Action readiness transforms a neutral world into one with places of danger and openings towards safety, in fear, with targets for kissing and their being accessible for it, in enamoration, with roads stretching out endlessly before one, in fatigue, misery, and despair, with insistent calls for entry or participation or consumption, in enjoyment. (Frijda 2007, p. 205, quoted in Deonna and Teroni 2012, p. 80)

Now, the obvious question is what exactly 'action readiness' is. I will mention two options and opt for the second.

The first option is that 'action readiness' is just a state of the body that is ideal for performing a certain kind of action. For instance, fear might just put the body in a state that makes one run as fast as possible. These states would be well correlated with what one tends to do in situations like this (perhaps because of evolutionary reasons), but they *would not* have a direct *causal* contribution to action. The motivation to flee would be independent of the emotion of fear and the emotion would just come in aid of this motivation.

This view would render the connection between emotions and action quite thin. Moreover, the role of emotions would be somewhat secondary: they would indeed typically aid in performing actions, but what one decides would be independent of emotions.

The second option, the one I favour, is that emotions already indicate courses of action and, importantly, they already constitute a *motivation* towards those courses of action, not a potential motivation, but an actual one[29]. Of course, this motivation can be defeated by competing motivations and it is not my purpose in this thesis to give a full account of how decisions to act are formed and what factors can come into play. My thesis is only that the emotion is already in a way part of the will, that it pushes towards some courses of action.

The main reason for claiming that emotions already constitute a form of motivation is that it fits better with the theory of attachments that I have argued for in the previous chapter. Indeed, I have argued that emotions are (at least sometimes) manifestations of attachments, which are drives encompassing what the agent cares about. The emotions capture the relevance of the situation to someone having these drives. And this relevance is what leads the agent to act.

This option also captures better the phenomenology of emotion described above. The courses of action indicated by the emotions are not experienced as some kind of possibilities that have something to be said for them, but as courses of action that the agent is already motivated to pursue (even if she eventually does something else, based on another motivation).

[29] I get the impression that Deonna and Teroni also verge towards this option, but they don't explicitly endorse it. This might be partly due to their efforts to distinguish clearly emotions from desires (Deonna and Teroni 2012).

There is one potential problem with the option I have proposed. It seems that sometimes there are no actions associated with an instance of emotion. One example is that of engagement with novels, in which the emotions felt by reader towards characters seem not to lead to any actions and indeed cannot lead to any of the standard actions associated with those emotions. I don't want to commit to a view on this problem, but it seems to me that these emotions do lead, for instance, to concentrating on various aspects of the novel, pondering on what a certain character might have done, reading further to find out what happens and so on. Someone might worry that some of these might not be strictly speaking actions, but irrespective of how we characterise them, they involve, at least sometimes, the will. My point is that mental or behavioural changes associated with an emotion, whether towards real objects or towards fictional ones, are motivated by that emotion, not merely correlated with it.

I would like to emphasise that an emotion need not indicate only one possible course of action. In some cases, various courses of action are just ways of pursuing one goal, so in a sense they are not really different courses of action: in fear, for instance, flying the scene or freezing are both options that aim at one's survival. But this doesn't seem to be always the case. In shame, for instance, one might be tempted to hide one's defects, but one might also be tempted to repair them. The first course of action is a way of not being seen by others, while the second need not be so and can be taken even if one knows that no one else will be aware of the defect one is ashamed of again. Or, to take an even better example, when one sees good friends that one hasn't seen for a while, the felt joy might lead one to lose oneself in a conversation with them or to step back and only regard them, sipping every word they say. What happens in these cases is that the emotion pushes us towards both courses of actions and we end up performing the one with the stronger push[30].

We can now also say something about what I listed as the fifth desideratum for an account of emotions, namely that emotions are triggered

[30] This tacitly assumes a picture of action according to which action is the result of competing 'pushes', that could be seen as vector-like, pointing towards various courses of action. Of course, in order to really argue for such a picture, there are many details left to be filled in, which is beyond the purpose of this thesis.

to some extent automatically, such as when one feels fear before judging that ‘there is a bear approaching’ or joy when bumping into an old friend before judging that ‘this is my friend, whom I haven’t seen for three years’. This phenomenon is usually seen as evidence for the epistemic role of emotions: they would attune us to values quicker than our slow reasoning process (Robinson 2005). In light of what I have said so far, I think this need not amount to anything more than the fact that we often realise what significance a situation has for us before analysing it carefully. Of course, it is an interesting question why we have this reaction so quickly, and this fact might have various scientific explanations, but it is not relevant for my purposes here what the explanation is.

Moreover, emotions clearly help us navigate situations involving a lot of information. For example, if there are many animals around us, a bear and one hundred pheasants, the fear of the bear will make any other potential emotion, such as curiosity about the pheasants, disappear. The emotion thus focuses our attention and makes us leave aside aspects of the situation that are not relevant to that emotion. But this focus is the result of what the emotions are manifestations of, including attachments.

6. ATTACHMENTS, EMOTIONS AND INTELLIGIBILITY

In this last section, I want to round off the discussion of emotions by explaining how attachments manifest themselves in emotions that push the agent towards various courses of actions and how these emotions render the courses of actions intelligible.

The attachment that X has for Y is a standing mental state that encapsulates the importance that Y has for X. This attachment leads X to react to various situations emotionally. As per attitudinalism, X reacts by her body taking an attitude towards the situation. She starts seeing the situation in light of this emotion and her attention is concentrated towards the salient aspects of the situation as indicated by the emotion. Moreover, the emotion pushes her towards various courses of action. For example, in the case of

Bernard Williams facing the situation of his wife and a stranger being in the peril of drowning, his attachment to his wife would presumably create a kind of fear in him, not a paralysing fear, but a fear that would make his body as alert as possible and would indicate saving his wife as *the thing to do*. His heart would race and his muscles would almost push him towards his wife, but he need not be, and probably would not be, aware of the individual physiological changes that happen, but only of how these changes lead him to view the situation. Moreover, everything around, including the stranger, would go completely out of focus, to the point of not receiving any attention.

We can return to Troy Jollimore's (2011) take on Williams' example, discussed in the last chapter. He claims that it is constitutive of love that in some situations, such as Williams', it silences other reasons. So he might agree with my description of the reaction of a loving husband, but he would also claim that this reaction is constitutive of love. Here, I disagree. Given the fact that the attachment is what causes the emotion that focuses Williams' attention, yet is distinct from that emotion, we can now see that this is not constitutive of love, but indicative of love. Indeed, if the attachment that constitutes the man's love for his wife is indeed strong, then it will manifest itself in an emotion that would interpret the situation in light of the peril of the wife drowning and will direct all the attention towards that. Perhaps it is a very good indication of love, in the sense that whether a husband loves his wife is extremely well correlated with responding as described above, but still, it is not constitutive of it.

So, emotions are a very important step between attachments and action, in that attachments indicate courses of action in emotions. But I want to claim that they are more than an intermediary between attachments and actions. I want to claim that in emotions, the courses of action that they indicate appear as what can be called, for want of a better term, 'intelligible'. It is in virtue of emotions that the mental life of an agent 'makes sense from the inside'. Let me try to detail this claim.

We sometimes say of an action that it is 'intelligible' if we know what end it serves. If we see someone running on the street wearing business clothes, we might find this bizarre, until we find out that they are late for work, in which case their running becomes intelligible as a means to getting to work on time.

However, this is not all there is to rendering human actions intelligible, as even if the end of an action is transparent, one could have chosen another end. Now, let's try to imagine what a life without emotions would be like. The agent would presumably experience various ends as having a kind of 'pull' towards them and decide towards that that pulls hardest and pursue it, but these ends would appear to come out of nowhere. When deciding to get away from something, there would not be a way to distinguish whether this action is done out of shame, contempt or fear – it would just be a decision to get away. The agent would not make much sense of her own decisions. It is only in emotions that the ends that 'pull' towards them are intelligible *from the agent's point of view*.

It can be replied that some ends can appear as rational or valuable in themselves without the need of any emotion and that makes them intelligible. Perhaps this is true, but surely not all ends appear as such. In particular, those that are set by out attachments, such as talking to Y, whom we have an attachment to, rather than to Z, cannot appear as such, as we do not expect other people to have the same ends – perhaps other people are attached to Z. So without emotions these ends would indeed come out of nowhere.

Moreover, it is not merely that in emotions various courses of action are intelligible as 'pulls'. Rather, the agent experiences themself pulling in those directions. For this reason, I think it would capture better the phenomenology to say that the agent 'pushes' towards a certain course of action in emotions, rather than that one course of action pulls towards it. Again, I think the attitudinalist theory is better placed to capture the phenomenology than the perceptualist one. Indeed, the latter agrees that various courses of action appear intelligible in light of an emotion, but this is because the agent perceives a certain value in an object and that value 'calls for a response'. In the attitudinalist theory, the change in how the world is experienced stems from the mode that the agent directs at the content, and even though we might say metaphorically that a situation 'calls for a response', it is clear to the agent that this 'call for a response' is created by her, by the emotional mode she directs at the content.

I want also to quickly return to the idea mentioned in the previous section, that one emotion can indicate two or more potentially different or even incompatible courses of action, that the same episode of shame can push one both to hiding oneself and to remedy one's defect, while the joy of seeing one's friends might push one both towards losing oneself in conversation with them and to stepping back and beholding them. It is important to observe that the emotion one experiences groups these two motivations together, it renders them intelligible as both indicated by the same emotion. It is transparent, I take it, to the agent that feels joy at seeing her friends that both courses of action stem from the same emotion of joy. If there were just the pulls towards the two courses of action, the agent would not experience them in any way as being grouped together, as stemming from the same source.

There is a worry directed at what I've said so far. Someone might think that there are emotions that are not experienced by the agent as oneself pulling in a direction and in which the courses of action they indicate are not rendered intelligible. Examples might include recalcitrant emotions or 'disowned emotions'. Examples of recalcitrant emotions include fears derived from phobias: people might fear spiders while knowing that they won't hurt them. 'Disowned emotions' are emotions that seem to be disowned by a purported 'higher self': for instance, we might have strong pangs of envy, while at the same time feeling that we'd rather not have them, pangs that we later feel very ashamed of and claim that 'we don't know how to get rid of them', as if they are some sort of disease. One might claim that all these emotions, or at least some of them, do indeed pull towards some courses of action, but the pull is not experienced as reflecting the self – it is experienced as an exterior force that one wants to get rid of, not as oneself pushing, as it were.

In reply to this, I will say essentially that in the case of disowned emotions, this is wishful thinking and not true. Just because one would rather not have an emotion, this doesn't mean that it is not the result of what one cares about. Rather, the interpretation that the emotion is not representative of oneself is the result of a kind of meta-emotion, directed at a first-order emotion. If I care about my conforming to an ideal of a non-envious person, this might lead me to regret or even resent experiencing envy towards X. This might even lead me to think, à la Harry Frankfurt (1971), that my envy is not really representative for me. But this is wishful thinking: perhaps my caring about not being envious is related to a stronger drive than the one that leads to my envy for X and the first drive 'wins' in forming the general thought that 'envy is not representative for me', but this doesn't mean that the less powerful drive and the envy it generates are not also mine, representative of me, etc. The meaning that the drive gives to the situation in terms of envy is genuinely representative of what I care about.

In the case of recalcitrant emotions, it is true that we do not like that we have those emotions and that they interfere with our lives. But this doesn't mean that they do not afford some kind of intelligibility, in the sense discussed above, of the actions they indicate. Isn't it the case that sometimes actions done from phobias look bizarre to the observer and only the person experiencing the emotion finds them intelligible, sees the actions, such as running away from the spider, as the natural thing to do?

To sum up, emotions render actions intelligible for the agent and allow her to experience them as a manifestation of what she cares about. This does not mean that she understands her attachments. In order to understand her attachments, she needs to experience many emotions and to put them together. But experiencing emotions and reflecting on them is the starting point.

Importantly, it follows that if we are to try to understand a person, we need to put together their emotions in order to conjecture what attachments they have. But in order to understand their emotions, it is not enough to understand the courses of action that they lead her to pursue. Given that these courses of action are intelligible to her in light of the phenomenology of her emotions, we need some access to the phenomenology of her emotions. In other words, we need to imagine her emotions. It is this that I will discuss in the next chapter.

7. CONCLUSION

I have argued that emotions are best conceived as bodily attitudes directed at a content given by another mental state. The phenomenology of an emotion is a function both of the attitude, which consists in bodily feelings, and the content, which is ‘coloured’ by the emotion. Emotions indicate courses of action and motivate the agent towards them. Moreover, they serve to make these courses of action intelligible to the agent in such a way that we can say that she experiences herself as pushing in those directions.

CHAPTER 3. IMAGINING EMOTIONS

1. INTRODUCTION

Having given a theory of what emotions are, how they relate to attachments and why they are important in the life of a person, we can now naturally ask how we can imagine the emotions of other people. This is something we do in our daily lives and it is essential for understanding other people – just think of invitations like 'Try to imagine what Jane must be feeling at this moment!' It is also essential, as we shall see in the following chapters, to understanding fictional characters and their attachments.

In this chapter, I will defend a theory of what it is to imagine an emotion of another person, what mental state imagining an emotion is. I will argue that to imagine an emotion is to form what I will call 'a thick meta-representation' of that emotion. It is a representation of that emotion and, given that emotions are themselves representational, it is a meta-representation. It is thick in the sense that one can grasp the phenomenological properties of the object represented, in this case an emotion.

The plan of the chapter is as follows: in the second section, I will work on framing the question, giving many examples of instances of imagining emotions. This is important because, as we shall see, we need to distinguish the phenomenon we are interested in from other related phenomena, with which it can be easily confused. I will then highlight, in the third section, four desiderata that a theory of imagining emotions should fulfil. In the fourth section, I will argue against the simulationist account of imagining emotions, while in the fifth section, I will defend the meta-representational account. I will end by saying a few words about the phenomenon of perspective-shifting in the sixth section.

2. THE PHENOMENON OF IMAGINING EMOTIONS

We often try to imagine the emotions of other people. They experience the world in a certain way in their emotions, with a certain phenomenology, and by imagining their emotions, we try to grasp this phenomenology without having the emotion ourselves. If a friend is angry about a remark that seems innocuous to us, we might try to imagine their emotion in order to understand how they see that remark and so to make sense of their reaction. We might not succeed and hence have to accept as given that the remark angered them. But if we succeed, we form a mental state of imagining their emotion, which allows us to understand that emotion 'from the inside'. It is this mental state, that allows us to grasp the phenomenology of another's emotion, to get some grip on how the world appears to them in their emotion, without having the emotion ourselves, that I am interested in. I don't want to give an account of what the best way to arrive at a good imagination of an emotion is, but just to describe the mental state of imagining an emotion.

Putting together the words 'imagination' and 'emotion' might lead one to think about the emotions we have towards fiction. For instance, we feel sad about the fate of Anna Karenina. However, in such a case, we are not imagining emotions, but having emotions towards a content that we imagine: the emotion is sadness and the content, namely, that 'Anna Karenina had such-and-such a fate', is imagined. As discussed in the previous chapter, we can have emotions towards contents borrowed from various mental states, including propositional imagination. Even though we imagine emotions when reading fiction, for example the emotions of fictional characters, the phenomenon of imagining emotions is not primarily related to fiction, so to avoid confusion I will focus mostly on imagining emotions of real people in this chapter.

The description of imagining emotions that I gave might lead people to think of the word 'empathy', but I will avoid using this word, because it is used in too many ways (see Matravers 2017 for a survey) and can lead to confusion. Still, we need to distinguish imagining emotions from various other phenomena that are sometimes labelled empathy. One is the phenomenon of putting ourselves in someone else's shoes. In doing this, we imagine being in someone else's situation and reacting emotionally to what happens to them. For instance, seeing my neighbour being insulted, I might imagine being insulted in the same way myself and, if I imagine the situation vividly, I might get angry. Again, as in the previous case, even though we imagine a situation, we have a genuine emotion towards that situation, rather than imagining one. Moreover, the emotion I feel might be very different from the emotion my neighbour felt – perhaps she never gets angry.

Another phenomenon that is labelled 'empathy' is a kind of perception of emotion in another person's body (Zahavi 2008, 2010; Zahavi and Rochat 2015). For instance, one might see the joy in another's face in a very direct way, without inferring that the other is joyful from their exterior appearance. I find it plausible that this phenomenon and similar ones exist and it is an interesting question what kind of mental states these involve, yet not the one addressed in this chapter. I only aim to give an account of the mental state of imagining the emotion 'from the inside', that would involve a glimpse of how the emoter herself experiences that emotion.

I also want to stress that I am primarily interested in the specific mental state of imagining an emotion, which is distinct from what is sometimes called 'perspective shifting' (e.g. Goldie 2011a). The latter would involve imagining a larger scale process of thought of another person, encompassing not only individual mental states, but their succession, their causing one another and so on. I will discuss perspective-shifting in the last section, showing how my theory could helpfully reframe some of the debates in this area.

Now, providing a theory of what it is to imagine an emotion depends on the theory of emotions one starts from. In the previous chapter, I defended the attitudinalist theory (Deonna and Teroni 2012), according to which an emotion is a bodily attitude directed at a content. The phenomenology of the emotion depends, according to this account, both on the content and on the mode (or attitude). The theory of imagining emotions that I defend in this chapter will work for the attitudinalist theory but also for other theories. Essentially, the theories of emotions to which what I will say in this chapter applies share two assumptions. First, emotions are intentional mental states, directed at various objects in the world. If one fears a dog or that the stock market might crash, one's fear is directed at the dog or at the potential crash of the stock market, respectively. Second, emotions change how the world appears or, to use a common metaphor, emotions 'colour the world'. If one fears a dog, one doesn't experience fear and at the same time perceives the dog as if one weren't feeling fear. Instead, the dog appears as threatening, dangerous or fearsome, and one's seeing, hearing, thinking of it are modified by the emotion. Various aspects of the dog, such as its pointed teeth, its little fixing eyes and so on, combine together to create a vivid impression of fearsomeness[31].

The question we are tying to answer is thus essentially the following: how can we imagine experiencing the world as the person having the emotion does? We can now see that once we have phrased the question like this, it seems conceptually puzzling, even paradoxical, how we can imagine emotions without having them: if one doesn't fear puppies, it's not at all obvious what would take to approximate in one's mind seeing puppies as fearsome. Bringing to mind an image of a puppy and then seeing it as fearsome seems to amount *not* to imagining fearing the puppy, but to *actually* fearing it. This is because we can experience fear of an object that we don't currently perceive, but only have a mental image of. Besides the initial conceptual difficulty, it is also difficult in practice to imagine certain emotions like fear of puppies – 'They look like the most innocent creatures, I can't imagine how you can see them as fearsome.' We'll get there.

At this point, I should also say that I am only giving an account of imagining emotions and not of imagination altogether. I am not taking a stance on whether all the mental phenomena that we call 'imagination' are unified, or whether it's just an unfortunate confusion that we call very different phenomena 'imagination'. Even if they can be unified, I think it is

[31] The two other theories of emotions discussed in the previous chapter, those of Tappolet (2016) and Roberts (2003), also share these two assumptions. Some of the things I say in this chapter might also work for other, more body-focussed, theories of emotions, such as those of Jesse Prinz (2004), but I take it that the question of imagining emotions is less interesting for those theories.

methodologically sound to investigate only imagining emotions. Moreover, some of the desiderata that I will discuss and some of the arguments that I will make do not work in the case of, say, visualising. Even though there are parallels to this phenomenon that I will mention throughout the chapter, my arguments do not aim to show that, for instance, visualising amounts to imagining seeing (Martin 2002), which would be an analogous thesis to mine.

To show the importance of the mental state of imagining emotions, I will discuss various ways in which we employ this mental state. These are part of various projects that an agent might have, and all of them involve imagining an emotion. This discussion will also help us distinguish cases in which we imagine emotions from superficially similar cases in which we do not imagine emotions.

First, we can use imagination of emotions to predict other people's behaviour[32]. If we want to predict how a person will act, a common trick is to imagine what they feel now and follow their mental process in imagination, ending up imagining the decision they make. What we do is essentially to follow 'from the inside' the causal process of mental states that leads to the action. For instance, if we see someone angry, we might imagine how their anger leads them to concentrate only on the object of their anger, how they feel a kind of pressure coming from the inside and how they unleash the pressure by shouting. Hence, we might predict that their anger will lead them to shout.

We can contrast this project with the following one. Suppose that instead of trying to imagine the emotion the other person feels, we imagine us being in their situation, in their external circumstances and then react emotionally to that scenario. We might, for instance, imagine what we would do if we were subject to the same offence. Bringing the possibility vividly to our mind, we might feel some anger and feel the need to shout. Importantly, whereas in the first project we imagined an emotion, in this one, we don't imagine an emotion, but rather *have* an emotion.

[32] This is part of the simulationist hypothesis (e.g. Gordon 1986; Ripstein 1987; Heal 2003; Goldman 2006), though usually the simulationists do not refer specifically to emotions. I am not claiming and usually neither are simulationists, that we *need* to use imagination this case.

However, imagining emotions is not always a means to an end. Sometimes it is an end in itself, and here we have the second way we use the mental state. We often want to approximate how it feels to have the emotion that another person has right now, to approximate the phenomenology of that emotion, without any further end in mind[33]. I use the term 'approximate' because it is not a very committed term and captures what I am looking for. The idea is that we are interested in the experience of other people, in their emotions as they experience them, irrespective of their causal effects in the world. We sometimes want to understand what someone is feeling even if this will not change anything, an interest that can be the result of a certain attraction to them, of us wanting to bring them closer to us. Just as we might want to be physically close to them, we want in a way to be mentally close to them. This is a simple, yet profound fact.

Third, just as we imagine the emotions of other people, we can imagine the emotions of fictional characters. Indeed, this is central to our engagement with many realist novels that attempt to depict in great detail the mental lives of the characters and attune us to subtle but highly revealing nuances. There are obvious questions about imagining fictional characters' emotions – do we attribute them to a real entity?; is there a right way to imagine them? – but irrespective of the answers to these questions, it seems clear that in practice this is how we read and for some novels this is the only way we *could* meaningfully engage with them[34].

Fourth, we can imagine the emotions that we have experienced in the past. For instance, I might imagine how I felt at my last birthday. We should note that some instances of bringing to mind emotions that we've had in the past might be claimed to be episodes of a different mental state, namely remembering, and there are good reasons to think that this is indeed a distinct mental state (Teroni 2017). Even if this is so, we can still imagine our past emotions without experiencing them as 'remembered'. If, for

[33] I am not assuming that the emotion is something 'private' that can only be accessed by the emoter. Indeed, as mentioned above, there could be a sense in which we see the joy in someone's face. However, there is a particular way in which the emoter experiences that emotion and this is what I am interested in.

[34] It is true though, as Matthew Kieran argues, that sometimes attempting to imagine the emotions of a character might be detrimental to our engagement with a novel, for instance when the character is narrow-minded and this narrow-mindedness can be more easily understood 'from the outside' (2003b, pp. 71-3).

instance, there is no trace of a past emotion left in us or if we just can't bring it to consciousness, we can try to reconstruct that emotion based on what we know of ourselves and in this way essentially treat our past self just as we treat other people[35].

Fifth, we can imagine not only our past emotions, but the emotions that we would have in the future if we were to find ourselves in certain circumstances. In a scenario that is not so rare in the modern world, suppose that I am considering moving to the countryside and commuting to the city for work. In order to decide whether this is a good idea, I should try to see how I would feel if I make the move. I could do this in two ways and distinguishing the two ways will help us sharpen up our understanding of what it means to imagine an emotion.

For one, I could vividly imagine the context in which I would find myself, for instance seeing the same people every day on the platform without having ever talked to them or waking up and seeing the sun rising in the distance. If I immerse myself in these scenarios, I could have some emotional reactions that give me a decent clue as to how I would react were the circumstances actual. Importantly, this does *not* amount to *imagining* emotions. Indeed, even if I react to an imagined scenario that involves some kind of immersion and some mental images, I would have *actual* emotions.

This is not the only way to project myself into this possible future scenario. I might assume that if I end up moving to the countryside, I will slowly change in such a way as not to have the emotional reactions that I have right now, or I might just be unable to work myself into an emotional state at this moment. In these cases, it makes little sense to apply the scenario above and it would be clearly better to *imagine* the emotions I would have when waiting on the platform for the morning train. If I imagine the emotions I might have, I do not need to have any emotions right now (Goldie 2012, pp. 81-3). By imagining my potential emotions, I essentially treat my future self from a third-person perspective[36].

There is also a sixth way we use the imagination of emotions. To use an example of Peter Goldie's (2000, p. 204), faced with a well-preserved

[35] Moreover, perhaps a similar analysis to the one I will provide for imagining emotions might be given for remembering them, but this is beyond the scope of this chapter.

[36] It is, of course, open to debate which strategy might be better.

Roman road, I might wonder what a Roman must have felt when walking down this road. I could then try to imagine an emotion felt by a Roman. This is ambiguous, as I might want to imagine either something like what the *average* Roman felt or, less ambitiously, an emotion that *some* Roman might have felt, irrespective of how idiosyncratic it might be. Similarly, I might wonder how it felt to be in London during the Blitz or how it felt to be a child in the Middle Ages. All these projects are inescapably vague, but they clearly involve imagining emotions.

3. DESIDERATA FOR A THEORY

Having highlighted the importance of imagining emotions, let's proceed to analysing this mental state. I start by discussing some intuitions about imagining emotions, which I will treat as desiderata that a theory should account for.

First, and perhaps most importantly, imagining an emotion is a different mental state from having that emotion. Even more, the two states have distinct phenomenologies: in the case of the imagination of emotion, the imaginer does not experience the imagined emotion as hers, so we might say, still in intuitive language, that there seems to be some distance between her and the imagined emotion. Yet at the same time, we can say that the imaginer gains some access to, or 'gets a grip on', the phenomenology of the imagined emotion, that after imagining it, she has an idea how that emotion feels[37]. So, there has to be some kind of relation between the state of imagining an emotion and the state of having that emotion, which would allow the imaginer to get an idea of the phenomenology. A theory of imagining emotions should thus primarily be able to account both for the difference and for the uneasy sense of similarity between having an emotion and imagining that emotion.

The second point is that, in some cases (though not always), when imagining an emotion of someone else, the attribution of that emotion is not

[37] The imaginer gains this access whether or not the imagined emotion is actually experienced by a particular person. If she tries to imagine the emotion of a particular person, but imagines it incorrectly, she still gains some access to the phenomenology of *an* emotion, just a different one from the emotion had by that particular person.

a further step from that act of imagination. By this, I mean that we don't first imagine it and then think 'This is actually X's emotion'. Rather, we imagine the emotion *as* X's emotion, so a theory of imagination should ideally be able to incorporate the attribution in the mental state of imagination. This is not to say that we necessarily attribute all the imagined emotions – I could just as well try to imagine how someone *could* feel fear in a certain situation without thinking that anyone in particular does feel fear in such a situation.

Third, we sometimes imagine someone's emotion in a very detached, almost clinical way, while other times we feel something like the emotion ourselves, we 'join in'. Usually, the first kind of imagination happens when we are not personally invested and have some other kind of interest, perhaps intellectual. We are just curious about what the other person feels, without being invested in what they feel. The second kind can happen, for instance, when we care about the other person, as when we imagine the sadness of a loved one and become sad ourselves. This does not mean that we just become sad and forget about the imagined emotion. We are aware all the time that we are *also* imagining the emotion of someone else.

These two phenomena, of detached episodes of imagination and of episodes in which we join in, are actually just the simplest cases of a broader class of phenomenona. What happens is that sometimes, when we imagine the emotion of another person, we infuse the imagination with our own emotions. Suppose a mother imagines the determination of her son to get the best mark in his class in a History exam. She may at the same time be proud that he is so determined and yet feel that he is a bit naive and can't realise that this mark is not as important in the grand scheme of things as he currently thinks it is. One might think that these attitudes are judgments that are separate from her imagination of her son's determination and that the imagination is just *causing* them. I think a better description of what happens is that her imagination is *infused* with these judgments – actually, these 'judgments' are more like emotions of the mother. She imagines her son's determination *as* admirable yet a bit naive. Another similar case is discussed by Goldie (2012, pp. 38-9): one evening, he has one beer too many, stands up on the table and starts singing in front of his friends, gleefully losing himself in moment. Next day, whenever he remembers the event, he is so ashamed that he can only imagine his glee as coloured by his current shame. To sum up, imagining can be infused with the imaginer's emotions in such a way that the phenomenology of the imagination is modified.

The fourth and last point is a technical, but important one. The broad idea is that there are degrees of precision in the imagination of emotions. When we imagine emotions in our daily life, we often do so *with a degree of vagueness*. Even if we know that X's fear of a dog and Y's fear of a dog are different and have different phenomenologies, we might imagine them in roughly the same way, indeed as we imagine most people's fear of dogs. We might know that for X the centre of fearsomeness is in the teeth, while for Y in the claws, but this need not make it into our episode of imagination. On the other hand, there are also moments – for instance when imagining emotions of people we know very well or when reading a very good piece of fiction – when the emotion we imagine is so precise, so thoroughly individuated that we feel that we have never imagined the same emotion before. There are two ways to cash out this difference and I want to insist on the distinction between them.

One way is to say that, in the first type of cases, we imagine a generic emotion. We can compare this with the following situation: when asked to think of a house, we might just think of the most common house, a model that has been reproduced over and over again, and not of a house with character. Similarly, when we are to imagine an emotion, say of fear, we might just imagine the most common version of that emotion. Even if we know that people's fears of dogs have very different phenomenologies, we sometimes don't bother with this and just imagine what we take to be the most common fear of dogs.

The second way to cash out this phenomenon is to claim that, in the first type of cases, we imagine an emotion *schematically*, without all the details, so essentially we don't imagine a complete emotion[38]. In the case of

[38] This is a similar idea to that of the speckled hen in the philosophy of perception (see e.g. Tye 2009). There, the thought is that even if you see a speckled hen, you might not see it as having an exact number of speckles. I am not taking a stance as to how this problem works in the philosophy of perception.

an actual episode of fear of a dog, its teeth are in some way part of the phenomenology of the emotion – either they play a central part in the appearance of fearfulness of the dog or they do not. If we imagine that episode of fear schematically, we do not represent the teeth as either playing a central role in the phenomenology or not playing a central role. We just leave this open, as it were. To return to the comparison with thinking of a house, we might think or form a mental image of a house while leaving it open whether it is made of brick or of stone. We should now note, and I will return to this, that no house in itself can be schematic – its walls, for instance, are made of something, so in reality it cannot just be left open what they are made of. Only a *representation* of a house can be schematic in leaving some details unfilled.

To see why the second option is more plausible, let's think of the following scenario: suppose that you are imagining X's anger at Y; this anger might include a certain interpretation of Y's behaviour – for instance, X might think something like 'After what he did to me, he is now even having a laugh with his friends' and hence, as part of the anger, see Y's laughter as obscene. Now, not knowing this about X's construal of Y's laughter, when imagining X's anger, do you imagine one of (a) Y's laughter definitely being part of the anger in the way described above and (b) Y's laughter definitely not being part of the anger in the way described above? If you imagine neither of the two, that means that you leave it open whether Y's laughter is interpreted in that way in the anger and hence imagine the anger schematically.

To sum up this section, the four desiderata are: (1) imagining an emotion is a different mental state from having that emotion, with a different phenomenology, but it also gives the imaginer an understanding of the phenomenology of the emotion; (2) sometimes, we attribute the imagined emotion to someone as part of the imagination, not separately from it; (3) some episodes of imagination are coloured by the imaginer's perspective; (4) some episodes of imagination do not involve all the rich details of the imagined emotion; this can be interpreted either as imagining generic emotions or as imagining emotions schematically, with the latter option more plausible.

4. THE SIMULATIONIST VIEW

I will now present the simulationist theory of imagining emotions and argue that it fares rather badly with respect to the desiderata highlighted above. Importantly, I am only arguing against a simulationist theory of imagining emotions, not against analogous theories applied to other mental states, such as beliefs or perceptions. I will actually explain where my objections apply specifically to the case of emotions.

The broad idea of this theory is that imagining an emotion E consists in forming a simulated version of E, that is, a mental state that is similar to E in some respects, something like a copy of E, but is 'off-line', i.e. it does not result in any behaviour specific to that kind of emotion. For instance, if I have the mental state of fear of a dog, this might lead me to run away. If, on the other hand, I have the mental state of simulated fear, I do not have any such behavioural reaction (even if the dog is in front of me.)

When it comes to similarities between E and simulated E, things get trickier. Goldman, for instance, claims that '[a]t least three categories of resemblance are eligible: introspectible, functional and neural respects of resemblance' (2006, p. 49). He is more interested in questions of attributing mental states to other people and in prediction of people's behaviour, so he places very little emphasis on phenomenological similarity, claiming that anyway, 'cognitive scientists will place little credence in introspectible resemblances' (Goldman 2006, p. 49). But if the goal of imagining an emotion is not so much prediction or attribution, but some kind of understanding of the phenomenology of the emotion (the experience, as it were, 'from the inside'), the phenomenological differences and similarities between E and simulated E are essential. We need to explain how the state of simulated E gives the imaginer an idea of the phenomenology of E.

A more promising way of developing the simulationist theory comes from Currie and Ravenscroft (2002). They actually claim that there is imagination only of beliefs and desires and not of emotions[39], but we can

[39] In earlier work, Ravesncroft (1998) discusses emotions as well in the context of simulation, but he doesn't seem to make clear whether in the process of simulation, we have genuine emotions towards content given by simulated beliefs or simulated emotions.

easily extend their view to emotions as well. Moreover, even if other simulationists don't phrase their theories like Currie and Ravenscroft, I think that they have in mind something like what Currie and Ravenscroft say.

The assumption we are starting from is, as in the previous chapter, that we can analyse mental states, at least the intentional ones, in terms of mode (or attitude) and content. A mental state would consist in a mode directed at a content and many attitudes can be directed at the same content. For instance, one can believe, assume or hope that 'it will be sunny tomorrow'. Believing, assuming and hoping are modes, while 'it will be sunny tomorrow' is the content. I will use the word 'content' quite broadly, to encompass everything that mental states can be directed at, whether propositional or not: for instance, if one fears a dog, the dog is the 'content' of fear.

Now, the simulationist's claim would be that imagining an emotion E consists in having a mental state that has a similar mode to E and is directed at the same content as E. For instance, if E is hope directed at the content 'that it will be sunny tomorrow', imagining E would consist in forming a mental state whose mode is similar to hope, that we could call 'hope-like imagination', directed at the content 'that it will be sunny tomorrow'[40].

Now, let's think how we could cash out the difference between the two modes, of hope, on one hand, and of hope-like imagination, on the other. Of course, as discussed before, one difference is that one leads to behaviour and the other does not. But we are also interested in the phenomenology. One option would be to claim that the two modes give rise to the same phenomenology, but, as mentioned before, this seems plainly wrong. If two mental states had the same phenomenology, they would feel the same to the agent – yet, it is clear that the agent knows from the phenomenology whether she is imagining an emotion or actually having it[41].

[40] I believe that this is very similar to, if not the same as, Walton's view of quasi-emotions (Walton 1990). He discusses this similarity in Walton (2015). At the risk of simplifying her view, I think this is also the position taken by Vendrell Ferran (2022).

[41] Or at least in some cases. In the case of visual perception, in some cases the agent sees something but mistakenly thinks that she actually imagines that thing (Perky 1910), so it is not implausible that this happens in some cases of imagining emotions. But these are marginal cases and in most cases one can easily tell the difference.

So the simulationist should assume that the imaginative version of an emotion gives rise to a different phenomenology. However, given that it consists in a similar mode directed at the same content, the difference should not be a significant one, just one whose main role is to make it clear for the agent that she imagines an emotion and does not actually have it.

One way the simulationsit could go from here is to take inspiration from Sartre and claim that the simulated version of the emotion involves some kind of absence that is manifest in the phenomenology. In his discussion of the difference between perception and forming a mental image, Sartre argues that the latter, unlike the former, posits its objects as absent (Sartre 2004 [1940], pp. 11-14). Sticking to the mode-content language, this could be expressed as follows: the mode involved in forming a mental image involves, amongst other things, taking the content as absent. This positing of content as absent explains why the difference between the mode of perception and the mode involved in forming a mental image is not one of degree, with one being a weaker version of the other, but one of kind. It would also explain the phenomenological difference between perception and forming a mental image. Following this idea, the simulationist might try to explain the difference, including the phenomenological difference between E and simulated E, in a similar way, with the simulated version involving some kind of absence.

I don't think this strategy can work. For Sartre, in the case of mental imagery, it is the content that is represented as absent. In the case of emotions, however, the difference between having an emotion and imagining an emotion has nothing to do with whether the content is in some way present or absent. Indeed, one can feel an emotion towards something that is not present or even towards something fictional, as when we fear a character from a book. Conversely, one can imagine an emotion directed at something that is present, as when one sees a dog and tries to imagine the fear that other people feel towards that dog, a fear that one does not in any degree feel. The difference between mental imagery and imagining an emotion is that if we are to talk about some kind of absence, in the case of imagining an emotion the key absence that we are interested in is the absence of the mode of emotion, not the absence of its content. And, given that according to the simulationist theory (as described here) the emotion is only simulated and not represented in imagining it, that is, it is not part of the content of the imagination, it cannot be represented as absent.

We can thus see that the simulationist has trouble explaining the difference in phenomenology between having an emotion and imagining it. Of course, she could claim that the phenomenological difference is *sui generis* and cannot be described – you know it when you see it. This wouldn't be the worst move, as clearly not all phenomenological differences can be put into words (try to describe the difference between seeing red and seeing green), but given that we talk about a difference between different *kinds* of mental states, we should prefer a theory that is able to say more on this matter.

Having discussed how the simulationist theory fares with respect to the first desideratum discussed above – i.e. about the difference between having an emotion and imagining it – we should move on to the next three. For the purpose of discussing them, I will assume that simulating an emotion has a very similar phenomenology to having that emotion, yet one that is not identical, with the difference manifesting the fact that it is an episode of imagination and not one of having an emotion. A simulated emotion would be a copy of that emotion that somehow manifests itself as a copy.

Regarding attribution, remember that when we imagine an emotion of a specific person, the attribution of the imagined emotion to that person is not an ulterior mental state ('The emotion that I have just imagined is X's), but is concomitant with the act of imagination. Even more, we might want to say that we imagine the emotion *as X's*. Can the simulationist theory at least partly account for this?

One option the simulationist might go for is to say that besides the mental state of simulated E, I also have a belief state that 'This is what X is experiencing', where 'this' refers to simulated E. This belief would accompany the experience of simulated E. I think this strategy has something to be said for it, but it faces two main problems. First, the imaginer cannot have the belief that 'This is what X is experiencing' for the simple reason that she knows that what she experiences is *different* from the actual emotion – we've just discussed that there is a difference in phenomenology. So, what she would have to believe is more like 'This is the simulated version of what X is experiencing, but it's close enough'. This starts to look ad-hoc and clumsy, and it's not at all obvious that this is what we have in mind when we are imagining an emotion of someone else. What seems to go wrong is that the phenomenological difference between having E and imagining E should not create a problem for the attribution of the emotion to another person. Quite the contrary: the phenomenological difference seems to account for the fact that we do not experience the emotion as ours, but as the person's whom we attribute it to. Rather than creating clumsiness, the phenomenological difference should fit in well with the attribution of the emotion to someone else.

The second problem with this strategy is that the simulated state and the attribution seem to be too separated. To see this, we might consider a case in which I experience an emotion and believe that the person next to me is experiencing the same emotion. I might then have the occurrent belief that 'This is what the person next to me is experiencing' and we would have a kind of attribution similar to the one in the case of imagination. Yet, as explained above, in the case of imagination we want something more than this: we want the emotion to be imagined *as* X's, not to have another incidental belief that attributes a similar mental state to X.

The complications faced by the simulationist are the result of the fact that, given the way we have built the theory, it is hard to make sense of how the imaginer can imagine an emotion *as* X's. It doesn't seem possible to be part of a mental state that it (the mental state) is attributed to someone else. I experience all my mental states as mine, not as someone else's, even if someone else happens to have similar mental states. The simulated emotion is experienced by the agent as her own and the fact that a non-simulated version is attributed to someone else necessarily involves some further mental process. This sort of problem does not appear in the case of mental imagery, because one can attribute the mental image ('this is in Paris') at the level of content. In the case of a simulated emotion, it is hard because the attribution would have to happen at the level of mode.

The third desideratum was related to the fact that our imagination of an emotion of another person might be coloured by our own emotions, e.g., a mother can imagine her child's determination with pride. How could the simulationist account for this? We should note that simulated E and the imaginer's emotion E1 that colours E should not just mix up, as when someone has a love-hate emotion or a fear-excitement one. In the case in which one's emotion is colouring a simulated emotion, there should be a hierarchy, with E1 not just mixing up with simulated E, but being directed at it. To see how this could work, let's compare this with case of one emotion of an agent being directed at another (non-simulated) emotion of the same agent. I might, for instance, be amused by something slightly immoral and, at the same time, feel shame directed at my amusement. This would be different from the case of fear-excitement, in which the two emotions are directed at the same object. Indeed, shame would be a second-order emotion directed at the first-order emotion of amusement. The amusement would then be coloured by the shame. Perhaps something similar happens in the case of simulating an emotion E. We might have an emotion E1 that would be like a second-order emotion, like shame in the previous case, but this time directed at simulated E. The second-order emotion E1 would then colour the simulated emotion E.

This solution works to some extent, but still, it faces one significant problem. According to the picture described above, the emotion E1 is directed at the mental state 'simulated E', that is at one of the imaginer's mental states. But this sounds wrong, for the emotion E1 should be directed at the emotion E, experienced by the target. If I imagine my friend's amusement and feel shame about this amusement, my shame is not directed at my imagining of the amusement. It is directed at my friend's (non-simulated) amusement. Even more, the hypothesis that the imaginer's emotion E1 is directed at simulated E, and hence only indirectly at the emotion E of the target, doesn't seem to capture the phenomenology – the imaginer's emotions is experienced directly as being directed at the target's emotion E, not indirectly, by being directed at 'simulated E'.

We finally get to the fourth desideratum, namely that there are degrees of precision in imagination, that we can imagine an emotion with all the rich phenomenology or not. I have argued above that there are two ways to interpret our imagining an emotion without all the rich phenomenology: either we imagine a *generic* emotion, that is, one that does not have an idiosyncratic phenomenology, or we imagine the emotion *schematically*, that is, with some details left out. Now, I think the simulationist theory is bound to take up the first option. Indeed, given that simulated E is a similar kind of mental state as E, it is not obvious how it could be schematic. An instance of fear without an idiosyncratic phenomenology is not a schematic instance of fear (what would that be?), but a generic fear, one very similar to the fear experienced by other people. If the teeth of the dog one fears are not in any special way manifest in the emotion, that just means that they are not relevant for the phenomenology of the emotion. Similarly, if simulated fear does not involve seeing the dog's teeth in any special way, this does not mean that it leaves open how the dog's teeth are experienced in the actual fear that one simulates, but just that they are not experienced in any special way. This is because, as we said, the simulated emotion is a copy of that emotion, so is a similar kind of mental state. Only representations can be schematic, as when we draw a house with only a few lines; copies cannot be.

In conclusion, the simulationist view struggles to account for many aspects of imagining emotions. What seems to be the problem is that according to this theory, the mental state of imagining an emotion is *too similar* to the mental state of the emotion itself. Of course, the claim that the imagination is similar to the emotion itself is driven by the fact that imagination is supposed to give some access to the phenomenology of the emotion, but we should try to find an alternative theory that can still account for this access.

5. THE META-REPRESENTATIONAL VIEW

As might already be clear by now, the view that I am proposing will locate the imagined emotion at the level of the content of the imaginative mental state. Given that an emotion is a representation, imagining an emotion would be a meta-representation. The tricky part will be to build my theory in such a way as to show how we can get some kind of understanding or access to the phenomenology of the imagined emotion in the imagination.

The idea of locating a mental state to be imagined at the level of content is not new, but as far as I know it has been applied mostly to sensory imagination, by people claiming, for instance, that to visualise an object is to imagine the mental state of seeing it, by which they mean that in this episode of imagination, the content is a mental state of visual perception (Martin 2002; cf. Peacocke 1985; Gregory 2016 for a recent overview of the debate that ensued)[42]. When it comes to emotions, I have only found one attempt to apply this idea, that of Fabian Dorsch (2012, pp. 337-364). His account is rather short and based on a purported analogy between imagining emotions and imagining pain[43]. He discusses the case of pain and then says that the same might be applied to the case of emotions. In the case of pain, he explains the connection between the phenomenology of being in pain and imagining pain, that is, forming an imaginative mental state with the content involving the mental state of pain, in the following way:

> The idea is that, while a feeling of pain involves painfulness by instantiating it, the imaginative (or mnemonic) awareness of such a feeling involves painfulness by representing it as instantiated. As a result, feeling pain and imagining it are subjectively similar in that both their phenomenal characters involve the quality of painfulness. But they differ from our first-personal perspective in that they involve this qualitative aspect

[42] Here is Mike Martin's Dependency thesis: 'to imagine sensorily a φ is to imagine seeing a φ' (2002, p. 404). He does say something about imagining itches, but it is not clear to what extent he thinks his view could be extended to other mental states. His not going into this is understandable, as he essentially builds his view to serve as an arbiter between representational and disjunctivist views of perception.

[43] I am not even sure that Dorsch is trying to answer the same question as me. It seems that what he is really concerned with is the puzzle of our emotional reactions towards fictional events, and argues for his view of what he calls 'emotional imagination' in terms of its ability to solve this puzzle. He claims that what happens in that case is the following: we have a quasi-emotion (Walton 1990) towards what happens in the fiction, and then we imagine having that quasi-emotion as a real emotion in the fictional world, and it is the latter that counts as emotional imagining. He doesn't say that having a quasi-emotion is necessary for emotional imagination, but he doesn't say otherwise either. Anyway, as I argued above, I think it is a mistake to associate imagining emotions only with fictional or hypothetical scenarios.

> in different ways: the former is really an experience of pain, while the latter is an episode of representing pain. (Dorsch 2012, p. 362)

Leaving aside the obvious worry that feeling pain and having an emotion might not be interchangeable, I take it that more needs to be said about how the representation of an experience inherits the phenomenal character of that experience. Even more, I've suggested above that it can be misleading to say that having an emotion and imagining that emotion have similar phenomenologies – rather, we should say that the latter helps us gain some access to the phenomenology of the former. This is what a theory needs to explain.

My starting idea, inspired by Richard Wollheim (2015 [1980], esp. pp. 137-51), is to distinguish between what I will call, for want of better terms, *thick* and *thin* representations, a distinction that will apply to both mental and other kinds of representation. Let's start with some examples. A cat can be represented by a drawing and a photograph, on the one hand, or by a word, such as 'cat' or her name, 'Izzy', on the other. In the first case, the cat is *apprehended in* the representation, while in the second case, she is merely referred to. As a second example, we can think of signs on public lavatories: to indicate that a lavatory is for women, one can either write 'Women' on the door or draw a woman, the former representation being thin, while the latter thick. Moving on to mental states, one might have a visual perception of a cat or one might have a belief like 'Izzy the cat is tired'. In the case of visual perception, the cat is apprehended *in* the perception, while in the case of belief she might be merely referred to. If we try to give a definition of what it means to 'apprehend' the object in a representation, the best thing would be to say that this consists in the representation allowing the agent to grasp in some direct way some of the qualitative features of the represented object from the representation. Now, I want to call *thick* representations those in which the represented object is apprehended in the representation and *thin* representations those in which it is merely referred to. It is part of the phenomenology of the former that the object is experienced in some way as part of the representation.

For the sake of clarity, I want to distinguish these notions, of thick and thin representations, from two other related notions. First, although the idea of apprehending something in a representation is similar to, and actually an expansion of, Richard Wollheim's (2015, [1980], esp. pp.137-151) idea of *seeing* something *in* a painting, they are not quite the same. Wollheim develops his notion in order to explain how people, facts, etc., can be pictorially represented in a medium that is very distinct from them, namely paint on canvas, and the notion itself is related to apprehending something *in a medium*. In my case, I definitely do not want to claim that mental representations involve some kind of medium, such as sense-data, so my notion is not the same as Wollheim's.

Second, the distinction that I want to make between thick and thin representations is not the same as the distinction that Jerry Fodor (2008) makes between iconic and discursive representations. For him, a discursive representation is one which has a '*canonical* decomposition into parts' (Fodor 2008, p. 173). In the case of a sentence, the parts are the words – a sentence cannot be decomposed by being interrupted in the middle of a word or even of a letter. On the other hand, an iconic representation does not have this canonical decomposition – in a picture of a person, 'each picture part pictures a person part' (p. 173). As Fodor himself says, his definition implies that a graph indicating say the position of a particle in time would be an iconic representation. But according to my definition, this is not so, as no object is apprehended in the graph.

Now, I want to argue that just as there are simple thick representations, as in the examples above, there are also thick meta-representations, that is thick representations of representations. In the case of paintings, this is obvious: we can think of a painting of a painting. In the case of mental representations, this is a bit less intuitive, but there is no reason to assume it shouldn't be possible.

So, here is my proposal: imagining an emotion E consists in forming a thick meta-representation of the emotion E. Unlike the simulationist theorist, I propose that the content of this mental state is not the content of E, but E itself, and the mode is some kind of imagination specific to emotions[44]. Also, this meta-representation is thick, in that the imaginer can apprehend E in it, and so this imaginative meta-representation is different from other thin meta-representations, such as the attribution of an emotion in a belief ('She is experiencing fear of the dog'). To use a metaphor, while the simulationist claims that imagining an emotion is forming a kind of *copy* of that emotion, that is, a similar mental state, I would claim that an episode of imagination is *a painting of an emotion*[45].

Now, we shall see how the introduction of this extra degree of thick representation can nicely account for the four desiderata. Even more, I hope that I will account for them in a way that seems intuitively right.

First, let us consider the difference between having an emotion and imagining an emotion. The two mental states have distinct contents: while the emotion E has a certain content, imagining E has 'E' as a content. They also have distinct modes: while E has an emotional mode, imagining E has an imaginative mode. Besides accounting for the difference between E and imagining E, by introducing an extra degree of representation, my theory is very well placed to account for the difference and similarity in phenomenology. Indeed, this extra degree of representation introduces a certain *distance* that explains why that the imaginer does not experiences the emotion as hers. Yet the representation being thick accounts for the fact that the imaginer gains some access to how the emotion is experienced, in that she apprehends the imagined emotion in the mental state of imagination. Coming back to the analogy with painting, the difference between having an emotion and imagining an emotion is a bit like the difference between seeing a scene and seeing a painting of that scene. A painting is a very different kind of object from the scene itself, yet it can give a good idea of how the scene looks. Similarly, imagining an emotion is a very different mental state from having the emotion, yet it can give a fairly good idea of the phenomenology of the emotion. And here is my key conclusion: it is *not*

[44] I don't want to claim that the mode involved in imagining emotions is *the same* as in other kinds of imagination, for instance, in imagining propositions as licensed by a novel. It might just be that we use the word imagination for mental states that have very little in common.

[45] Dorsch uses a similar metaphor when giving his account of visual imagination (2012, p. 333).

necessary to have a *similar* mental state to the emotion in order to have access to its phenomenology.

Regarding attribution, the imagined E is located at the level of content, so it can be attributed at the level of content. Just as we can form a mental image of something and attribute that image to a particular place (e.g. Paris) at the level of content, so we can imagine an emotion E as the emotion of someone in particular.

The third desideratum was related to the fact that some episodes of imagination can be detached, while others can be *coloured by* or *infused with* the imaginer's emotions. To account for this, we'll use the following general idea: that emotions colour the content of other mental states. If we are afraid of a dog that is in front of us, our perception is coloured by the emotion of fear. Or, to take another example, we might form a mental image of a traitor and feel contempt for him, in which case the mental image is coloured by the emotion of contempt – that is, we form the image of the traitor as the object of contempt. In the case of the imagining an emotion, the content of the mental state is the imagined emotion, so that content can be coloured in a similar way by an emotion we might feel. Indeed, when I have emotions towards the emotions that I imagine – say I feel disgusted by someone's amusement – I imagine the relevant emotion, i.e. the amusement, as the object of my disgust in a similar way in which I visually perceive the dog as the object of my fear.

Now, in some cases, the emotion we feel towards the imagined emotion of another is a kind of 'doubling' that emotion, of 'joining in'. If a friend is happy that she has published a paper, I will also be happy when I imagine her happiness directed at the published paper. If she is sad after a loss, I will imagine her sadness with sadness. I said in the second section that I shall avoid using the word 'empathy', but I take it that here we have an interesting phenomenon that is empathy-like, at least in the common, non-philosophical language. Someone might worry that this is not quite empathy-like because our emotion, as I described it, focuses too much on the other person, on her emotion, and not on its object. If we are empathetic, the thought might go, we should feel sad about her loss, not about her sadness. I will use this potential worry to clarify the phenomenon further.

When I am imagining my friend's sadness with sadness, it is true strictly speaking that my sadness is directed at her sadness, but we need to remember that her sadness is directed at her loss, so her loss is part of her sadness. The objection seems to construe the imagined emotion as a mere fact about the other person, not taking into account the essential intentionality of the emotion and the way this intentionality plays a role in the mental state of imagining that emotion. Again, a comparison is appropriate: when I am in awe of a painting or find it graceful, it is true that my emotion is not directed at the scene depicted, but we cannot say that the scene depicted is in no way connected to my emotion. My emotion is directed, in a way, at the scene *as depicted* in the painting. Similarly, in the case of imagining sadness with sadness, my sadness is directed at her loss *as the object of her sadness*.

Finally, let's consider the fourth desideratum, that one can imagine an emotion with more or less detail. I said that there are two ways to cash this out: either we claim that if we imagine an emotion without many details, we actually imagine the *generic* emotion, that is, a very common version of that emotion, or we claim that we imagine the emotion *schematically*, that is, without all the details filled in. My theory naturally leads to the conclusion that we imagine the emotion more or less schematically. Indeed, it is common to many kinds of representation that they can be schematic. A drawing is the most natural example: we can draw a house while leaving it open what colour it is, for instance by drawing only the main lines. Of course, when we move to mental states, this is bound to be more controversial, but it still seems plausible that some mental representations can be schematic in this way. As stated earlier, representations can be schematic, while copies cannot.

6. ON PERSPECTIVE-SHIFTING

I will end by suggesting how the view that I have proposed might serve to advance a related debate in the philosophy of mind, namely, the debate on perspective-shifting. The broad question is in which ways, and to what extent, we can switch from our perspective to the perspective of another person, where the concept of 'perspective' is left ambiguous. First of all, there is a distinction that goes back to Wollheim (1984, p. 79; cf. Goldie 2006) between imagining being someone else, say X, and imagining being in X's shoes. (Here, Wollheim and Goldie use the word 'imagining' perhaps in a more ambiguous way than how I've used it in this chapter.) The first kind of project, imagining being X, would involve bringing to mind something like a sequence of mental states that X would have, given her dispositions, values, attachments, traits, etc., that give rise to X's emotional reactions. The second kind of project, imagining being in X's shoes, would involve bringing to mind a sequence of mental states that the imaginer would have, were she to be in X's situation, given her dispositions, values etc. Let's take each option in turn, starting with the second.

Regarding in-his-shoes imagining, we can now make an important distinction in light of the theory that I have proposed. Suppose that I want to put myself in X's shoes. Our discussion so far in this chapter points to two possible options. For one, I could try to vividly bring to mind X's situation and then react emotionally to that, that is, to have *actual* emotions directed at imagined situations. The second thing I could try to do is to *imagine* the emotions that I would have in that kind of situation. In this second strategy, I would use the knowledge I have of myself and my usual reactions to imagine the emotions that I would have when faced with certain circumstances. It is important to distinguish the two projects not so much because one would be better than the other (I remain neutral on this point), but because they involve very different mental states – one involves actual emotions, while the other involves imagining emotions.

Let's move on to the harder question of imagining being X. We want a definition of this such that the question of whether we can imagine being X is meaningful. This means that we cannot just equate imagining being X with having X's sequence of mental states, as I cannot just give up my perceptions, beliefs about who I am, where I am now, etc. (Matravers 2017, pp. 36-37). The most natural definition would then be the following: imagining being X is imagining the sequence of X's mental states, including the emotions of X, in their succession, with one seeming to follow from the other. If X has sequence of mental states M1 – M2 – M3, the imaginer would have to have the sequence imagining M1 – imagining M2 – imagining M3. To what extent is this possible? Peter Goldie argues that this is rarely possible, and only when the target is very similar to the imaginer (2011a). His main argument is that X himself is motivated in an unreflective way by his general values, traits etc., a way in which we, the imaginer, cannot reproduce. For instance, if X is a compassionate person, he might immediately feel compassion when seeing another person in distress. The imaginer, if she is not a compassionate person, might be able to bring compassion to mind in this imagined circumstances, but she would have to make an effort, saying to herself something like 'X is a kind person, so, if I want to switch to his perspective, I have to imagine compassion in this circumstances.' Or, even if the imaginer does not have these very explicit thoughts, she needs to have some kind of mental state that draws her attention to the fact that she has to be compassionate. Imagining compassion doesn't just come naturally. However, X doesn't have this kind of thought, nor does he have the reflective distance – he just feels compassion, as it were, unreflectively. Hence, while the sequence of X's mental state is: 'seeing person in distress' – 'feeling compassion' – 'deciding to act', the sequence of the imaginer's mental state would be: imagining 'seeing person in distress' – thinking 'I need to keep in mind that X is compassionate' – imagining 'feeling compassion' – imagining 'deciding to act'. So, the imaginer just cannot reproduce the sequence of mental states of X without distorting them by having these reflective thoughts that X does not have, so her attempt to shift her perspective to X's is inescapably defective[46].

I think that this argument of Goldie's works only if we assume that imagining an emotion is simulating that emotion, that is having a mental state that is very similar to that emotion. If simulationism were true, the simulated state M2 that follows the simulated state M1 would naturally be the state that would follow M1 in the imaginer (without simulating). In other words, if M2 tends to follow M1 in the imaginer in the daily mental life of the imaginer, then when they simulate M1, they would then tend to simulate

[46] Of course, as Langkau (2021) observes, this would not imply that the imaginer cannot end up imagining the individual mental states of the target well.

M2. If the imaginer is not compassionate, seeing a person in distress would trigger at best some discomfort, so this is what she would naturally tend to simulate after simulating seeing someone in distress. If, however, imagining an emotion is, as I have argued, a thick meta-representation of that emotion, there is already a reflective distance built into the mental state. The imagined mental state is not experienced as one's own and so the mental states that one imagines in a sequence don't have to come one after the other *unreflectively*, as if it were one's own. The reflective distance that Goldie is worried about is just built into the very mental state of imagination and so the imaginer can, if she is a good imaginer, have the sequence of mental states imagining 'seeing person in distress' – imagining 'feeling compassion' – imagining 'deciding to act', even if she is not at all compassionate. Goldie's objection stems from the assumption that when trying to represent another person's emotions, we either have a third-person access to them or we try to reproduce them in our own mind almost as they appear in the target's mind, in which case of course the distortion described by him would appear, creating a mess. Once we have the option I have described, the problem vanishes.

7. CONCLUSION

I have argued that to imagine someone else's emotion is to form a thick meta-representation of that emotion. This extra layer of representation introduces a distance between the imaginer and the imagined emotion such that she doesn't experience this emotion as hers, and allows us to account for the four desiderata much better than the simulationist theory: the difference between having an emotion and imagining it is like the difference between interacting with an object and with a representation of it; and the fact that the imagined emotion is represented at the level of content makes it possible for it to be attributed to someone as part of the imagination, infused with the imaginer's own emotions and imagined more or less schematically. The theory also allows us to see how we might imagine the 'perspective' of someone else, that is the sequence of mental states they experience, without distorting it. Given all these perks, I hope the meta-representational view will appeal even to those who might be reluctant to accept the existence of such a complicated mental state.

CHAPTER 4. THE SUBJECTIVE KNOWLEDGE THEORY

1. INTRODUCTION

The broad goal of my thesis is to explain how literature can help us understand the attachments amongst fictional characters and to show why this is of ethical relevance. This follows a broader and seemingly widely shared intuition, namely, that when reading works of literature, we understand characters, and that this can constitute a form of learning. Yet this is also the idea behind what is commonly known as the 'subjective knowledge theory' (SKT), which claims, roughly, that by reading literature we can gain experiential knowledge, or knowledge 'what it is like', of various experiences (Walsh 1969; Wilson 1983; Kajtár 2016; Bailey 2023). In this chapter, I will explore this theory and offer a defence of a version of it, which will draw on the theory built in chapter 3; however, I will also show its limits in accounting for what we learn from literature.

The version of the SKT that I will argue for is the following: first, literature can offer us knowledge of what various emotions of fictional characters are like, and the main way we get this knowledge is by imagining these emotions, as opposed to having them ourselves; second, this knowledge is not necessarily inferior to that acquired by actually having those emotions ourselves, even though in practice it might often turn out to be inferior.

The plan is as follows: I will start by discussing the idea behind the SKT, as well various observations made by its proponents. I will then explore two options for developing it: either we only gain experiential knowledge by *having* emotions ourselves when reading, or we also gain it by *imagining* the emotions of characters. To adjudicate between the two, I will delve into debates about what experiential knowledge is and argue that, on the most plausible version of what it consists in, by imagining emotions we can get experiential knowledge which is just as good as the one derived from having emotions. I will end by discussing the limits of SKT, notably that having experiential knowledge of someone's emotions is not enough for understanding their attachments.

2. THE SUBJECTIVE KNOWLEDGE THEORY

The subjective knowledge theory (SKT), usually traced back to Dorothy Walsh (1969), associates a particular kind of knowledge to literature, namely subjective, phenomenal or, as I shall call it, experiential knowledge[47]. The broad idea is that there is a kind of knowledge that in real life we typically get by having experiences. To take Walsh's favourite example, someone who has been poor and lonely in a big city can usually claim to know what it is like to be poor and lonely in a big city, while their friend who has always been comfortably well-off might have no idea what that is like. Yet, Walsh claims, it is not necessary to have the experience in real life in order to know what it is like – one can just get this knowledge from reading a novel.

László Kajtár (2016), a recent proponent of the theory, gives a concrete example from Cormac McCarthy's novel *The Road.* This novel, Kajtár claims, shows us what it is like to fear death, an experience we might not have had, or of which we might have only had a very weak and diffuse version. The example strikes a chord: set in America after an ecological disaster that left few survivors, the novel follows the journey of a father and son trying to reach the shore – and to survive. They fear death not as something that might happen in an accident but as something that is already there, in the ever-present smog, in their cart which might break down at any point and in all the other survivors, the majority of whom have organised into bands that kill and eat each other. The short phrases contribute to the ominous atmosphere, and further attempts to over-describe one particular source of fear would have unjustifiably focused the reader's attention and created a sense of a more acute fear of death, as opposed to the over-arching fear that has become a constant part of the protagonists' lives, to the point of

[47] I will use the terms 'subjective knowledge', 'phenomenal knowledge' and 'experiential knowledge' interchangeably.

not being noticed as such. It is difficult to pick an illustrative passage, as the novel works by drawing us into the characters' perspectives very slowly, but the following quotation captures the general tone:

> They left the cart in the woods and crossed a railroad track and came down a steep bank through dead black ivy. He carried the pistol in his hand. Stay close, he said. He did. They moved through the streets like sappers. One block at a time. A faint smell of woodsmoke on the air. They waited in a store and watched the street but nothing moved. They went through the trash and rubble. Cabinet drawers pulled out into the floor, paper and bloated cardboard boxes. They found nothing. All the stores were rifled years ago, the glass mostly gone from the windows. Inside it was all but too dark to see. They climbed the ribbed steel stair of an escalator, the boy holding on to his hand. A few dusty suits hanging on a rack. They looked for shoes but there were none. They shuffled through the trash but there was nothing there of any use to them. When they came back he slipped the suit-coats from their hangers and shook them out and folded them across his arm. Let's go, he said. (McCarthy 2019 [2006], p. 83)

Following the father and son on their journey and carefully reading all these descriptions that give a glimpse of how the world appears to them, we seem to get some knowledge of what their experiences are like. It is the goal of a proponent of SKT to explain carefully what happens in the mind of the reader that accounts for gaining this knowledge.

Before exploring ways to develop the theory, it would be useful at this point to discuss a few claims that its proponents have made. First, Walsh contends that just having experience is not enough in order to know what that experience is like. Instead, one needs to have what she calls, perhaps confusingly, 'an experience', that is, experience that one attends to or is aware of:

> An experience, as life experience, is self-consciously recognized by the experiencer as his. An experience is not just awareness: it is awareness of awareness. Animals, no doubt, can be said to have experiences but only a being capable of self-consciousness can be said to have 'an experience'. (Walsh 1969, p. 84)

The people that do not have this second-order awareness of their experience, Walsh claims, do not acquire the relevant knowledge of what it is like. They merely *are*, e.g. poor and lonely in a big city.

I am rather unconvinced by this. It depends, of course, how one cashes out the second-order awareness, but it seems initially plausible that the experiences that can be described as 'in the grip of …', with the dots filled by 'fear', 'anger', etc., do not involve this second-order awareness, for the agent is completely focussed on the object of their fear, anger and so on, and has no attention to spare on the experience itself. Yet it is precisely these experiences about which one can most plausibly be said to know what they are like, to such an extent that one can later be haunted by them.

Even if we don't accept Walsh's claim, there is something appealing in the idea that knowledge of what an experience is like involves some second-order mental states, that just having the experience is not enough. Given that the object of knowledge is the experience, we need to have some mental state directed at it. I hope to eventually show that this thought is right.

Catherine Wilson (1983), another proponent of the theory, claims that we should distinguish between a deep and a shallow way of knowing what an experience is like. The deep way would necessarily involve a change in what one takes to be reasonable. Talking about Newland Archer in *The age of innocence*, she writes:

> [T]here is both a 'deep' way and a 'shallow' way of 'knowing what x is like'. A reader may understand 'what it is like' to be Newland Archer in the shallow sense, and his philosophical convictions may undergo no revision. But on the version of the theory we are now considering, if he understands what it is like

> to be Newland Archer in the 'strong sense', his philosophical conceptions will necessarily be affected. The theoretical resources of Walsh's theory give out at just this point, for the mundane examples which were used to fix the concept of 'knowing what x is like' are of no help in understanding the stronger concept. The person who knows what it is like to be poor and lonely in the big city has not necessarily been forced to alter his conceptions of poverty, or loneliness, or anonymity. (Wilson 1983, p. 494)

This would mean that even though we might think that we can expand our knowledge of what experiences are like without changing our conceptions in life, Wilson thinks that this is only the shallow way of learning. She is very brief in expanding on this, saying explicitly that the details are left to be filled in about what it means to posses a conception (or concept) of 'poverty', 'honour', etc. (p. 496), and doesn't give too many reasons for her claim. I get the impression that she is driven by the idea that in some cases of acquiring experiential knowledge, we should have our ethical views of various actions or ways of living changed by such knowledge – for instance, by coming to know what it is like to be poor and lonely in a big city, we might come to see the brusqueness of some people that are poor and lonely as appropriate, or at least understandable, rather than impolite. This might be true, but it is not enough to justify her distinction between deep and shallow way of having experiential knowledge. First, it seems implausible that in all cases of experiential knowledge we should revise our conceptions – sometimes, we might think that the character about whose experience we learn what it is like has a mistaken world-view and that the knowledge we gain serves just to understand them better, not to learn something of ethical relevance from them. Second, even in cases in which we should learn something of ethical relevance, this is plausibly a further step from gaining the knowledge of what an experience is like. Mixing them together does not help us understand what is going on. (Compare with the following: knowing how the stock market works and how it impacts society should presumably change what we think about stock market regulations, but it does not follow that someone who knows better than everyone else how the stock market works and yet couldn't care less about regulations only has a 'shallow' form of knowledge.)

Now, so far we've talked about 'experiences', but this is ambiguous. Often, when we talk about one experience, such as that of being poor and lonely in a big city, what we actually mean is a sequence of mental states that are somehow connected to each other, either by one following another, or by their having a common cause, or the same object. When one is poor and lonely in a big city, one is sometimes afraid or suspicious of passers-by, the well lit shops might appear defiant rather than inviting and so on. This is a very complex experience that one understands bit by bit. So, to make things easy, I will concentrate on the knowledge of what one particular emotion is like. For all intents and purposes, I will assume that an emotion, even though it might extend in time, is a mental state that can be understood by looking at a time-slice of it.

Note that, as argued in the second chapter, the phenomenology of an emotion depends on the mode (e.g. fear) and on the content (e.g. 'that the dog is approaching', or just 'the dog'), so even if one knows what a particular kind of fear is like, this does not amount to knowing what all kinds of fear are like. Therefore, the knowledge of what an emotion is like is about a particular emotion (e.g. this fear, with all its details), not about fear in general. It is true, and I will come back to this, that there is a sense in which if one knows what one instance of fear is like, one might have a clue of what other instances of fear are like, a clue that is inaccessible to someone who does not know what any instance of fear is like; but the fundamental question of knowing what emotions are like is about particular emotions.

I've said that we are interested in 'particular emotions' and what they are like. A clarification is in order here: even though I talk about a 'particular emotion', I still refer to a type of emotions, rather than a token. Such type of emotions, call it 'emotion E', is very narrow, and its tokens have a lot of phenomenological aspects in common, but it is still a type. Hence, even if we want to know what an actual emotion of someone is like, another person could theoretically have a phenomenologically indistinguishable emotion, so what we are interested in is what that emotion as a type is like. Also, one could technically know what an emotion that no one has ever had is like, in that case having this knowledge directed at a type with no existing tokens.

To conclude this introductory discussion, we are left with the following simplified set of questions: can literature give us knowledge of what a particular emotion E is like? If so, how does it do this? Is this knowledge just as good as the knowledge that we get from actually having E? My answer will be: yes, it does give us this knowledge by helping us to imagine E, and this knowledge can be just as good as the knowledge gained by actually having E ourselves.

In the following two sections, I will analyse two options for how literature might give us this knowledge: first, by helping us to have an emotion ourselves (Bailey 2023); second, by helping us to imagine the emotion (Kajtár 2016). I will argue for the latter option.

3. KNOWLEDGE VIA HAVING AN EMOTION

According to the first option, we get to know what emotion E is like by having E ourselves while reading a work of literature (Bailey 2023). How do we end up having E? Essentially, what happens is that we put ourselves in a character's shoes, that is, we imagine being in their situations and react emotionally to those situations. Note that although we imagine the situation, and in doing that we imagine other mental states, such as perceptions, beliefs and desires, the claim is that we actually experience, rather than imagine, emotions. These emotions might be directed, for instance, at something that is supposed rather than believed, but, as shown in the second chapter, this is possible. And if these emotions are new to us, in the sense that we haven't experienced them before, then we might learn what they are like.

The first reason why some of these emotions might be different from the emotions we have in real life is that in real life we do not encounter the situations encountered by fictional characters. Fiction helps us experience alternative scenarios. However, there is still a sense in which the emotions we have towards these alternative scenarios are characteristic of us. For instance, I take it that for many readers of *Madame Bovary* it is hard to get themselves to feel contempt for Emma Bovary's husband, an ultimately decent man. Putting themselves into Emma's shoes, they would have completely different reactions from hers, reactions that reflect their values – for instance, they might feel sympathy or pity. This emotion would be very different from the contempt that Emma feels. The range of emotions that we end up experiencing cannot thus be too wide and there will be emotions of people that are different from us (for instance, Emma Bovary) about which we do not learn what they are like. One wouldn't learn what it is like to feel contempt for a decent but unimaginative man that happens to be one's husband. Therefore, according to this version of the theory, even if literature can give us some experiential knowledge, this knowledge doesn't seem to be as extensive as the knowledge we can in fact gain seems to be.

There is a way to respond to this initial worry and develop the theory, following ideas of Susan Feagin (1996). In addition to what I've explained above, she contends that when we imagine another's situation, we can also have a change of what she calls 'sensitivity', which she defines as follows:

> A sensitivity is the psychological state or condition that makes it the case that one will have a particular kind of emotional or affective response to a certain sort of phenomenon, or situation, or to what elicits the response. (Feagin 1996, p. 74)

To unpack this, let's look at an example that she adduces: if we read a parody, we have certain expectations with regard to what the role of the sentences should be. Hence, we can react to a sentence completely differently from how we would have reacted had we seen the exact same sentence in a serious work. This is just because we got into the sensitivity that makes us read the sentences parodically and thus have the appropriate reaction of amusement and light-heartedness.

Now, Feagin claims that when reading a book, we might have a change of sensitivity from our usual sensitivity to one of the characters' sensitivities and in this way we can end up reacting to various events in the book as a character would have reacted. Just as we can get ourselves into the parody-reading sensitivity, we can also get ourselves in the Anna Karenina sensitivity and react to her troubles as she would.

In recent work, Olivia Bailey (2023) proposes, without quoting Feagin, a very similar view, using the word 'sensibility' instead of 'sensitivity':

> A sensibility shapes how the world looks to one in two respects. First, a sensibility governs one's patterns of attention: a really timorous person is always on the lookout for features of the world that could be construed as threatening. Her mind effortlessly fixes on shifting shadows, sharp edges, and glinting teeth – things that a braver soul might typically not even notice. And second, a sensibility governs which evaluative construals are triggered or invited by the lower-level properties or features one notices. (Bailey 2023, p. 222)

Again, the question is how a work of literature can change our sensibility such that we come to have emotions that we would not otherwise have. To this, Bailey responds that we have, or at least some of us have, multiple sensibilities, and the reason we usually associate one sensibility with one person is that each of us tends to have 'one native sensibility, or alternatively, one native coherent set of sensibilities' (Bailey 2023, p. 232). Here, by 'native' I take it she doesn't mean that we were born with it, but that it is our dominant one, which everyone associates with us; later on, Bailey uses the term 'home' sensibility to capture the same idea. For instance, a person might be very serious and not prone to laughing at trivial matters, and about such a person we can say that they have a serious sensibility. Yet, as Bailey observes, even such a person can, on certain occasions, such as a dinner with old friends, adopt a more light-hearted attitude, and a joke that would have seemed silly on another occasion now appears charmingly naive, an occasion for amusement and mirth. All this would suggest that they somehow have this other sensibility in them, hidden, yet awaiting to surface. And it is this that literature has to do to in order to get us to feel new emotions, to draw out from the depths and bring to the front various latent sensibilities that we have and that do not manifest themselves very often in our daily life. Of course, this possibility depends on our actually having these sensibilities, so Bailey concludes:

> Interestingly, if this is right, then one's capacity to receive fiction's gift of empathy is not just a function of whether one is imaginative, in the sense of being able to readily assemble representations in novel combinations. Being a teachable reader is also a matter of being a relatively un-rigid person, one with other 'voices' that can readily be drawn out. (Bailey 2023, p. 234)

My first worry about Bailey's view relates to the very idea of a 'sensibility' (or 'sensitivity'), which is supposed to do the work in showing that we can have emotional reactions that are significantly different from our usual ones. Let's try to discuss further what this might mean. For instance, I might be morose when I see 80% of my acquaintances and very happy when I see the remaining 20%, but this is easily explained by the fact that I like 20% and dislike 80% of them, and saying that I somehow have a morose 'home' sensibility and a cheerful hidden one does not seem to add much. Indeed, if we want the notion of sensibility to do some work, it needs to have some kind of explanatory power, to be an extra factor besides other factors such as which people I care about, what I like and so on. Indeed, Bailey herself says that a sensibility 'shapes how the world looks to a person' and 'governs one's pattern of attention'. However, I am wondering whether such a phenomenon exists, or, anyway, if it exists, whether it can have the influence that Bailey ascribes to it. It seems to me that our changes in emotional reactions, changes that might seem to an unknowing observer to be inexplicable and hence as ascribable to some temporary change in sensibility, can often be explained by two observations, that sometimes apply at once.

First, we need to notice that the object of our emotions can be more complex than we describe in casual speech. Let's take an example: if a friend tells me 'you are cruel' in a context in which we tease each other, I might be amused; if she tells me the same thing (even on the same tone, with the same facial expression and so on) without any such context, I might be worried. This is because in the first scenario, I simply interpret it as form of teasing, of playing the game and so on, while in the second, I assume that she means it. In the first scenario, if I started to believe that she actually means it, my emotional reaction would change. What happens in these scenarios is not so much that I have different emotions with the same object consisting in my friend's saying 'you are cruel'. Rather, the object of my emotion includes what I take to be my friend's intention and the social context. In the same way, we can make sense of Susan Feagin's example of responding in different ways to the same sentence in two different books, one a parody, the other a serious work. The object of my emotions in the two cases would be different, it would not be just a sentence but a sentence as part of a certain work. I would get a third kind of emotional reaction if I saw the sentence painted on my door.

Second, as argued in the first chapter, besides objects, emotions can also have sources, that is, they might be manifestations of various drives, including attachments. The amusement at my friend's joke, a joke that would otherwise irritate me, might be a manifestation of the attachment to my friend. Indeed, there is no reason why attachments cannot manifest themselves in amusement. So even if I react to an aspect of the joke and my amusement has that aspect as its object, the amusement can be to some extent a manifestation of the attachment itself. It follows that whether I am involved in the joke and amused by it can be significantly influenced by whether I have an attachment to the person telling the joke or not. My being amused by a joke told by my friend and irritated by the same joke told by a stranger would thus be no more mysterious than my being hopeful for my friend's job application and not for a stranger's. Therefore, there is no need to posit some special change of sensitivity to explain why I am amused in some instances and not in others.

It's true though that there seem to be some cases which cannot be easily explained either way. Bailey provides one such case:

> If I've been recently menaced by a stranger, creaks and rustles will show up for me as much more significant than they would if I were navigating the world via my usual bold outlook. (Bailey 2023, p. 232)

Another example is that of having had a tough time at work one day and becoming very easily irritated or angered – 'you're in an irritable mood', we tend to say in such a case (though not in the way philosophers tend to use the word 'mood'). A third example would be that of inebriation, which might make everything more amusing than usual. In the three cases, it seems that the agent is more prone to fear, anger and amusement, respectively. Moreover, the change seems not to be explainable in terms of some change in drives – in all cases, the agent seems to care about the same things as before. Neither can it be easily explained in terms of the object of emotion – it would be quite a stretch to say that what happened at work somehow made it into the object of one's anger about someone talking too loudly on the train.

However, what tends to happen in these cases is that our emotional reactions are exacerbated. We have emotions that are still characteristic of us, but we over-react, in the sense that we have stronger emotions than we would have had were we not in the new mood. So I'm not sure to what extent these cases can help support Bailey's proposal that we have dormant sensibilities that await to be awaken in order to have emotions that we would not have otherwise. Even if having a bad time at work might get me to feel anger towards a person on the train by whom I would normally be just mildly irritated, it's not clear how such a change of mood could make me feel contempt for Charles Bovary, for whom I would not normally feel any kind of contempt. Moreover, it is not obvious that literature changes our mood in the same way in which a glass of wine or a tensed meeting at work does.

Perhaps a better account of sensibility can be given that has more explanatory power, but as it stands I cannot see how this can be done. Our emotional reactions might vary to some extent, according to what I called above a change of mood, but even if this is true, the change seems to be one of degree, and the emotions that we experience still reflect our values, attachments and so on. I am thus sceptical of the claim that we can experience all sorts of emotions that we would not naturally experience.

Notwithstanding how optimistic we are about how much literature might stretch our reactions to include emotions we usually do not have, I think there is a more significant worry about this version of SKT. It seems fairly clear to me that when we read fiction, we do not have the emotions that the characters have. To take an example, in *Anna Karenina* the eponymous character is worried, or even terrified, that she might be separated from her child. What do we, as readers, feel when reading this? If we adopt the book's outlook, we might be worried as well, but we would not be worried that 'we might be separated from our child'. Instead, we would be worried that 'Anna might be separated from her child'. The content of our emotion would be different from the content of Anna's. This might sound a bit pedantic on my side – does the fact that the content is a bit different make a difference? Well, of course it does – the directness of the emotions that Anna has towards something that is the centre of her life cannot be matched by the perspective of the reader, a perspective which, even if emotional, is that of an observer.

Even more, Anna's emotions stem, as I showed in the first chapter, from her attachment to her child, which I argued is a drive, that is, a standing mental state, in this case directed at her child, that manifests itself in emotions. As a reader, I don't have an attachment to Anna's child and, given that I don't have any children, I don't have a similar attachment to anyone. So, how can I try to get myself to feel an emotion that is similar to Anna's? I might imagine the belief that he (her child) is my child and about to be separated from me, and I might visualise him. But this is not enough. As shown in the first chapter, the emotion that Anna has for her child is not generated by her belief that he is her child. Instead, that emotion is a manifestation of her attachment to her child. In the absence of an attachment, the belief that he is her son and about to be separated from her might not generate any emotion. Or, if she has some kind of general valuation of parent-child relationships, it might generate a different emotion. So why should an imagined belief (or some other similar state) that a child is my child and about to be separated from me, even coupled with some kind of visualisation, generate Anna's emotion? It might generate some emotion, but that emotion would be a manifestation of, for instance, my general values regarding parents and children, not an emotion that actual parents have towards their own children.

If I had a child, I could imagine, when reading about Anna and her child, a similar situation happening to me and my child in the real world. In this way, I might have an emotion that is in some way similar to Anna's. If this emotion is new, it can be said that I've gained some new experiential knowledge. However, the range of emotions that we can experience in this way is very limited: not only are they typical of us, but their objects are the things we care about in real life – in this case, the object of emotion would be *my child*. Therefore, we don't even get experiential knowledge of emotions of people that are similar to us, yet care about different individuals, let alone of people that are different from us.

Furthermore, such a strategy would instrumentalise literature to the point of caricature. A novel such as *Anna Karenina* is about the characters in the novel, in this case mostly Anna herself, with their attachments, values, temperaments and so on. It is not illegitimate to compare their reactions with our reactions in the real world, but the main focus is still on what happens in the story and how it is depicted. If, when reading Anna Karenina, I focus on imagining various scenarios about *my actual child*, I am a terrible reader. The strategy that I described, that of focusing on my own child, essentially says that for the purpose of my learning what an emotion is like, it doesn't quite matter how the fictional character reacts, but just how *I* would react when placing myself in a similar situation to hers, involving people I know in the real world.

To sum up, the first version of the SKT claims that we learn what various emotions are like by having emotions ourselves. To do this, we imagine being in the characters' situations and react emotionally to those imagined situations. My scepticism towards this strategy is based on two main worries. First, I don't think that the range of emotions that we can get ourselves to experience is very wide and I doubt that it includes emotions that do not reflect our values, concerns, attachments and so on; for instance, it wouldn't include the emotions of people that are different from us. Second, when we read literature, our emotions are not those of the characters, but about the characters. As readers who get emotional about what happens in novels and who 'identify with' or, to put it a bit vulgarly, 'root for' characters, we can easily forget that our perspective is different from those of the characters and that our emotions are different from theirs. Thus, if we want to claim that we get some knowledge of what *their* emotions are like, we need to find an alternative account.

4. KNOWLEDGE VIA IMAGINING AN EMOTION

The second option is that we only imagine the emotions of characters, without experiencing them, and this is the main way we get knowledge of what those emotions are like. I take it that this is roughly the view that Kajtár (2016) adopts. Kajtár construes imagining an emotion as simulating that emotion, where by a simulated emotion E we mean a state that has a similar phenomenology to E yet different functional properties: it can be caused by mental states other than beliefs, such as suppositions, and it does not issue in behaviour. Now, in the previous chapter, I have argued that imagining an emotion is not simulating that emotion. Instead, imagining an emotion E is forming a thick meta-representation of that emotion. The fact that it is a meta-representation, so a representation of that emotion, implies that the phenomenology is different – there is a distance from the emotion, the emotion is apprehended as not being one's own. Yet the fact that it is a *thick* meta-representation implies that one grasps the phenomenology of the emotion E in the imagination of E.

Now, literature can help us imagine emotions of characters (and that of the narrator as well, but I will focus on characters). To take Kajtár's helpful example, Cormac McCarthy's novel *The Road* can make the reader imagine fear of death. As mentioned before, the fear of death as embodied by the two characters, father and son, colours almost everything around them: the ever-present smog, the deserted city, the people that they encounter. Note again that the reader might also feel fear, but not fear of death in the way the characters experience it. Rather, the reader would probably experience fear for the two characters, fear that something might happen to them, and so the object of fear would be very different: for the characters, the object is death; for the reader, it is 'that the characters might die'.

We can even note that if the novel had had the primary goal of causing fear in the reader and no interest in making the reader understand the emotion that the characters feel, it might have employed, as many lesser works of fiction do, all sorts of other techniques that cause fear. One such technique is that of dramatic irony, in which the work lets us see a danger that the characters face without being aware of, a technique that often gives rise to horror in the reader. The novel would read more like the following:

> As they were searching for food upstairs, the front door opened with a barely audible crack. 'Don't worry, it's the floor', he said to his son. Through the door entered a masked man, who scanned the room through his sooty little eyes. He took off his shoes. Through the holes of his noiseless socks, three bent nails were pushing out. As he made the first step on the stairs, he leaned forward and carefully drew a knife from his back pocket…

This caricature can at best cause some very instinctive fear and disgust in the reader, little different from the fear and disgust felt at some B-series films or novels. On the contrary, *The Road* does not use tropes that are known to cause fear in the reader and allows us to concentrate on imagining the perspective of the characters, creating a more serious engagement with their predicament. Imagining the emotions of the characters usually triggers emotions in the readers, but these emotions follow from understanding the characters and are often directed at the imagined emotions.

Now, before reading *The Road,* when asked to imagine fear of death, a person who hasn't had this experience might imagine very vaguely an emotion of fear that cannot be clearly labelled fear of death. After reading *The Road*, if they have a good enough memory, they might bring back to mind an imagined emotion from this novel, or a very similar one. This seems to suggest that our reader has learned, or at least improved her knowledge of, what it is like to fear death, or at least of one way it is like to fear death.

We can now return to Walsh's proposal that in order to know what a certain emotion is like, it is not enough to have that emotion; in addition to this, one needs to have attended to that emotion, that is, to have had a second-order state directed at that emotion. I have argued that that is incorrect, but I think there is something right in this idea, namely that knowing what an emotion is like is a matter of having some second-order state directed at that emotion. Indeed, imagining an emotion *is* forming a second-order state directed at that emotion. And the awareness that Walsh talks about is similar to the state of imagining that emotion, the difference being, of course, that in the case of imagination, the emotion is not present. Nevertheless, even if one learns what an emotion is like in Walsh's way, by having that emotion and attending to it, when thinking about it later, one imagines it, and so we might think of imagining the emotion as a kind of re-creation of the original awareness.

Now, perhaps there exists another emotion, besides what our reader has imagined, that can also be labelled 'fear of death'. If this is the case, can we say that our reader has learned what it is like to fear death? The problem with this question is that it is ambiguous: she has presumably learned what the fear of death E1, as depicted in *The Road*, is like, but perhaps not what another fear of death E2 is like. If all fears of death are to some extent similar and different from, say, fear of the crash of the stock market, then our reader might be said to have learned something (though not everything) about what all fears of death are like, namely that they are somewhat like E1. If fears of death are very different from each other, then presumably she hasn't really learned much about what a fear of death very different from E1 is like, but this is just to be expected. I will come back to all these issues in the next section.

One worry about what I have argued in this section is the following: how does the reader know that the emotion she has imagined is the emotion of the character? Perhaps she has imagined another emotion and wrongly believes it to be the emotion of the character. If that is the case, she hasn't learned what the emotion of the character is like, hasn't she?

There are various theories of whether there is a correct way to interpret a work of art, and those theories correspond to various ways to cash out which emotions, if any, the reader should imagine when engaging with such a work. However, I don't think this matters for what I am arguing here. Even if, as readers, we imagine the wrong emotion, that is, not the one that, according to some *right* interpretation, the character experiences, we still learn what a certain emotion is like, namely the one that we have imagined. In the case of *The Road*, if we imagine the characters' fear of death in a wrong way, it would be incorrect to say that we learn what *their* fear of death is like. However, we might still say that we have learned what *some* fear of death is like.

At this point, we've reached the main worry about gaining experiential knowledge via imagination, namely, that what I have described might not amount to knowledge. Surely, one might think, in order to know what E is like, you need to have actually experienced E. Against this, I will argue that we can get experiential knowledge via imagination and, moreover, that this knowledge can be, at least in some cases, just as good as the knowledge got via actually experiencing E. In order to argue for this, we need to have a clearer view of what experiential knowledge consists in.

5. EXPERIENTIAL KNOWLEDGE AND EVALUATING THE THEORY

I will analyse three options for what experiential knowledge consists in: propositional knowledge, knowledge-how and knowledge by acquaintance.

The first option is propositional knowledge. This hypothesis has been formulated by, amongst others, Frank Jackson (1986) in a completely different context, namely in the context of an argument against physicalism, known as 'the knowledge argument'. The set-up of the argument is the following: Mary is a scientist who has lived in a black-and-white room for all her life. She knows all the physical facts about colour vision, yet when she leaves her room and sees red for the first time, she learns something new, namely what it is like to see red. Jackson claims that what she learns is a proposition, yet he does not discuss options for what this proposition might be. Perhaps some options would be '*This* is what it is like to see red', where 'this' is a demonstrative that refers to her current experience, or 'Seeing red is different from seeing blue', or 'Seeing red is more similar to seeing orange than to seeing blue.'

This propositional knowledge is taken, in the original paper, to provide evidence that physicalism is false: Jackson claims that Mary learns a new fact and, given that she knew all the physical facts, this means that there are non-physical facts; therefore, physicalism is false. However, even if we grant that Mary learns a new proposition, it doesn't obviously follow that physicalism is false. Indeed, Tim Crane (2019) has insightfully argued that the argument could actually be not so much an argument in the debate between physicalism and dualism, as an argument about knowledge, with the conclusion not leading to either position. Starting from the premiss that Mary gains some propositional knowledge, the conclusion that Crane proposes is that there is some propositional knowledge that one can only gain by having certain experiences. This knowledge is knowledge of a fact, understood in the Fregean sense of 'a thought that is true'. Only someone that sees (or, as shall be explained later, imagines) red, can have the true thought 'This is what red looks like', but this does not mean that she has learned a non-physical fact. Essentially, this piece of knowledge is only available to people who have experienced red, because only they can refer to the experience in a direct way, and hence only they can have the thought that the knowledge is of. Debatable as Crane's interpretation might be, it suggests that we might discuss what Mary learns from having a new experience without necessarily taking a stance on whether physicalism is true, which is what I shall do in the rest of this chapter.

The second option (the 'ability hypothesis') is that experiential knowledge is a form of knowledge-how or, in other words, an ability (Nemirow 1980, 1990; Lewis 1990). This tends to be supported by physicalists and put forward as an alternative explanation to what Mary learns when leaving her black-and-white room. So, what might the object of such an ability be? For one, it might be the ability to imagine the mental state in question and to recognise the mental state when one has it. To know what it is like to see red would amount to, perhaps amongst other things, being able to imagine (correctly) seeing red and recognising a visual experience of redness as a visual experience of redness – 'Oh, so this car is red.' It would be this that Mary would have learned when she leaves her black-and-white room. Before, she might not have been able to imagine a visual experience of redness or to recognise a potential experience as one of seeing red. (However, it is not obvious that she couldn't have gained these abilities without leaving her room.) Similarly, coming back to emotions, having experiential knowledge of (a certain kind of) fear would amount to being able to imagine fear and to recognise an episode of fear in oneself as such.

However, Lewis (1990), in his account of the knowledge-how involved in experiential knowledge, lists, besides the ability to imagine and recognise the experience, the ability to *remember* it. Under the natural understanding of remembering, one cannot remember an experience if one hasn't had the experience, so by default one cannot know what an experience is like if one hasn't had the experience. Or, if one cannot remember but can imagine and recognise the experience, we could at best say that one has partial knowledge of what that experience is like.

I don’t think we should add ‘to remember’ in the list of abilities, besides ‘to imagine’ and ‘to recognise’. Here is one counter-example: suppose that I am afraid of flying and have experienced fear many times when the plane took off. If you now ask me if I can imagine an episode of fear of flying, I might be able to imagine it in great detail, with the engine seeming to stop and the front of the plane ready at any point to start pointing down. However, given that I have had the emotion so many times and that it has become, so to speak, routine, I might not be able to remember any particular instance of the emotion. According to Lewis’ version, in the case just described, I would thus not fully know what it’s like to have that emotion of fear. But this seems simply wrong: I would, perhaps more than anyone, know what that fear is like. The actual experience served to give me a very good knowledge of what the experience is like, and my being unable to remember one instance should not detract from this.

It might be replied that even if I cannot remember one particular episode, I can still, in some vague sense, remember the experience. But why should we assume this? We might suppose that all my episodes of bringing the emotion to mind are qualitatively similar and different from an episode of remembering. What would suggest that they involve some remembering, rather than mere imagination? These episodes of imagination might be accompanied by a feeling of familiarity with the experience, but this is clearly not enough for an episode to count as remembering. To see that just being accompanied by a feeling of familiarity is not enough for an episode of imagination to count as remembering, let’s think of an emotion that I have very often, say, impatience with the bus I have to wait for. When imagining this emotion, the episode of imagination would be accompanied by a feeling of familiarity in virtue of plain fact that this emotion is very much part of my life. But it is implausible to say that I cannot imagine this emotion without counting as remembering it, so the feeling of familiarity is not enough. Returning to the case of fear of flying, there is no reason to assume that I might not be able to imagine the emotion very well without being able to remember it.

An even stronger version of the counter-example, which should alleviate even the worries about remembering without remembering an exact episode, is the following: suppose that after I have many experiences of fear of flying, I stop flying, but I continue to imagine fear of flying from time to time. Doing this often enough, I retain the capacity to imagine it very well. Now, after giving up flying, I retain for a while the ability to remember the experience, yet it is plausible that after a long enough time, I cannot remember the experience any more, not even in a vague way, but I can still imagine it very well. Intuitively, it seems that in this case I know very well what fear of flying is like.

Another reason why we shouldn't take remembering as necessary for experiential knowledge is that, as argued in the second section, the object of this knowledge is a type of experience, not a token. Even if the type is very precise, involving many details (e.g. fear of death in a post-apocalyptical world, in which everything is seen as potentially contributing to death), it is still a type that can have many tokens – the emotion I know what it is like could, in principle, be had by anyone in an identical form. Given that the object of knowledge is a type, I think it's *ad hoc* to tie this knowledge to acquaintance with a particular instance.

Before moving on, I want to reply to an important objection to the ability hypothesis, in the version that does not include 'remembering'. The objection, raised by Earl Conee (1994), is that one can have the ability to imagine an experience without having experiential knowledge of that experience:

> Suppose that Martha is a superlative colour interpolator. She is highly skilled at visualizing an intermediary shade that she has not experienced between pairs of shades that she has experienced. Martha happens not to have any familiarity with the shade known as cherry red. She has seen, and vividly recalls, the look of burgundy red and the look of fire engine red. Suppose that Martha is now informed that there is a common shade of red, cherry red, which is a hue midway between burgundy red and fire engine red. At this moment, before Martha has imaginatively interpolated between those two shades, it is clear that Martha does not yet know what it is like to see

> something cherry red. She does not know this, although she is fully prepared to find out by exercising her imagination. Yet Martha already knows how to visualize cherry red, since she knows how to perform the imaginative interpolation between burgundy and fire engine red. Thus, knowing how to visualize something cherry red at will is not sufficient for knowing what it is like to see the colour. (Conee 1994, p. 138)

This is a significant worry, one that I will try to defuse carefully. What it claims is that we can have at the same time: an ability (or knowledge-how) to imagine an experience, in this case seeing cherry red; and no knowledge of what that experience is like. This would of course imply that an ability to imagine an experience is not enough for experiential knowledge. Essentially, the problem with this objection is that it conflates two conceptions of knowledge, or two approaches to knowledge. According to the first conception of knowledge, what we are interested in is what might be intuitively characterised as 'what we *really* know at this moment'; I shall argue that according to this conception, Martha has neither knowledge-how nor experiential knowledge. According to the second conception of knowledge, what we are interested in is what might be intuitively described as 'what we might not really know at this moment, but we can easily find out'; I shall argue that according to this conception, Martha has both knowledge-how and experiential knowledge. It will follow that the thought experiment poses no problem for the abilities view in the form I defend.

The first conception of knowledge is the one that Conee implicitly uses when he appeals to the intuition pump that Martha does *not* have the experiential knowledge of seeing cherry red. Indeed, it just seems that the experience of seeing cherry red is in some way distant from her, and she doesn't have the close relationship with this experience that we might have, for instance, with a belief that we have, which is the paradigm case of a candidate for knowledge. However, we should note that according to this same conception, it is not strictly speaking true that Martha has the ability (or knowledge-how) to imagine seeing cherry red. Indeed, before imagining seeing cherry red for the first time, she needs to interpolate between the two colours, burgundy and fire-engine red, to find, as it were, the colour that is half-way between them; therefore, there is an extra mental operation that she needs to perform, an operation that someone who can imagine seeing cherry red without interpolating has no need for. To see why this matters, we might think of cases in which a person can imagine an experience at the end of a long process. In one extreme case, I might say that I can imagine an episode of grief: to do that, I walk five metres to the bookshelf, pick a good book on grief, read it and get myself to imagine what grief is like in less than one hour. Of course, people would protest that this is cheating, that in order to count as having the ability to imagine grief, I should be able to do it without external help. But even if we stick to the ability to imagine grief using, as it were, just my mind, we can still modify the example to make it problematic: indeed, suppose that I know by heart the text of the book on grief, without having examined it at all – in other words, I've learned it parrot-fashion. I can then examine the text carefully in my mind and, after one hour, I manage to imagine grief. Presumably, even in such a case, I would not count as having the ability to imagine grief, but as having the ability to *learn* how to imagine grief. Returning to the case of our colour interpolator Martha, she also has to perform a mental operation in order to imagine cherry red; the fact that this operation takes a very short time and almost goes unnoticed creates the impression that she has the ability to imagine cherry red, but according to this strict conception of knowledge, the extra mental operation renders her as having only the ability to *learn* to imagine cherry red. She doesn't have the ability to imagine cherry red.

This phenomenon appears in other cases of abilities. Suppose that our friend Martha is a very good ice-skater, yet she has never used roller skates. Were she to try roller skates, she might be clumsy for the first few seconds but would very quickly, in a matter of minutes, become a good roller skater. She is therefore in a different position from someone who has never skated in any way, whom it would take way longer. For this reason, she might say, before trying roller skates, that 'of course, I have the ability to skate on them, it would just take a few minutes to get used to them'. This line might work very well to convey her state in common parlance, but we should note that, according to the conception of knowledge at hand, what she says is not true. Even though she can very easily get herself to skate well on roller skates, before she first tries them she doesn't have the ability to skate on roller skates. Rather, we should say that, in virtue of her ice-skating proficiency, she is able to acquire the other ability very quickly.

Yet we might think of a second, more lenient conception of knowledge. The intuition behind this is that there is a sense in which our colour interpolator Martha has the ability to imagine seeing cherry red more than someone who cannot imagine seeing cherry red no matter how many mental operations they perform. And it is this intuition that Conee seems to appeal to when claiming that Martha has the ability to imagine seeing cherry red. The conception of abilities (or knowledge-how) that we might end up with would encompass abilities that we don't strictly speaking have at this moment, but that we could easily acquire if need be. All this is similar to what happens in the case of beliefs. There, we might think that someone can believe that 'the prime factors of 91 are 13 and 7' even if, when asked, 'what are the prime factors of 91?', they would have to do a quick algorithm in their mind in order to come up with them. In trying to account for this phenomenon, we might extend our concept of belief to encompass propositions that we do not store in our head, but that we can easily come up with if asked (Stalnaker 1991). In this case, the extension of the concept of 'belief' naturally comes together with an extension of the concept of propositional knowledge, such that the person described above can not only believe, but know that 'the prime factors of 91 are 13 and 7'. Now, if we return to the case of experiential knowledge, we can think of the ability (or knowledge-how) to imagine as parallel to belief. Experiential knowledge would then be parallel to propositional knowledge. But this means that if we extend the concept of ability such that Martha counts as having the ability to imagine seeing cherry red, we should also extend the concept of experiential knowledge, to the effect that she also counts as having knowledge of what seeing cherry red is like. Under this more lenient conception of knowledge, this conclusion should not look suspicious, for we have already agreed that, under this conception, the object of knowledge is further away from one's current mental states than under the strict conception. To conclude this discussion, neither conception of knowledge poses a problem to the abilities view.

There is a third option for what experiential knowledge consists in, put forward by, amongst others, Conee (1994), Tye (2008), Giustina (2022) and Walsh herself (1969) in her book on SKT: that experiential knowledge is a special kind of knowledge that is *constituted by* acquaintance. In the case of experiential knowledge, the object of acquaintance is a mental state, but one can have knowledge by acquaintance of a person, or of a city – 'I know John' or 'I know London', in the sense that I have been acquainted with John or London (in a good enough way, to be specified, e.g. not by seeing them from the aeroplane). Here is Giustina:

> [S]imilarly to [Bertrand] Russell, I understand the notion of 'knowledge by acquaintance' not as knowledge *caused* by acquaintance, but as knowledge *constituted* by acquaintance: the kind of knowledge *of x* that consists in one's suitably direct awareness of *x*. … I argue that introspective knowledge by acquaintance is *sui generis*: it is a kind of knowledge that is irreducible to propositional knowledge. (Giustina 2022, p. 128, her emphasis)

One way to interpret this is that one has knowledge by acquaintance with an experience (or with an entity) only while one has that experience and attends to it. In this case, when one experiences jealousy, attending to it gives rise to this knowledge by acquaintance, but once the jealousy disappears, the knowledge disappears as well. I am happy to accept that there might be some interesting notion of knowledge by acquaintance along these lines, but I take it that this is not what interests us here. Indeed, according to this option, our experiential knowledge would be limited to the experiences that we have at present, so we couldn't in any way *accumulate* experiential knowledge, whether by reading or by having various experiences.

Another way to interpret this idea is the following: an agent has knowledge by acquaintance of experience x if and only if the agent *has* *experienced* (i.e. in the past) x and has attended to it in the right way (to be further specified). I think this interpretation cannot work. First of all, it ascribes knowledge to an agent at a certain time irrespective of the mental (or physical) states that the agent has at that time. Indeed, in the definition above, the only condition is to *have had* the mental state x, and no condition about the agent's current mental states appears. This is very problematic, since the idea of knowledge involves, even if it is not limited to, the relation that the agent has with the world at present. The second problem with this interpretation is that once we have knowledge by acquaintance of an emotion, we cannot lose it or, in other words, we cannot forget what the emotion is like. This is because, once we have the emotion, it will always be true of us that we have had it. But this is wrong: it should be possible to lose all kinds of knowledge.

We can think of a third interpretation: an agent has knowledge by acquaintance of experience x if and only if the agent has experienced x and has attended to it in the right way (to be further specified) *and* this experience is in some way retained. The worry regarding this interpretation is that it comes very close to the abilities view in the version that includes the ability to remember the experience. Indeed, whether the experience is retained in the mind seems to amount to whether the agent is able to remember it. And if she is able to remember it, she has surely had the experience. It follows that that this interpretation of knowledge by acquaintance is equivalent to a version of the abilities view. If the proponent of knowledge by acquaintance is to avoid the collapse into the abilities view, she has to give a different account of how the mental state that one is acquainted with 'stays' in the mind, whatever that might mean.

Walsh seems to think that one reason for cashing our experiential knowledge in this third kind of way is that the verification process is different from the process involved in propositional knowledge and knowledge-how. For propositional knowledge, we ask for evidence, while for knowledge-how, we ask for a demonstration by a display of the relevant ability. For experiential knowledge, however, it seems possible that some possessors of such knowledge can bring no other proof than the claim that they've had the experience:

> When someone says, with reference to some kind of human experience, “I know what it’s like. I’ve lived through it. I’ve experienced it,” we commonly accept that he does know, even when he cannot convey this knowledge. Knowing beyond saying is acceptable in such a case, not because saying is impossible, but because the only kind of saying that would be relevant is a saying that requires some degree of literary talent. (Walsh 1969, p. 104)

I agree that one might have experiential knowledge without having the power to convince others beyond giving one’s word, and very often for precisely the reason that Walsh invokes, but this doesn’t imply that the knowledge is not propositional or knowledge-how. Indeed, one might have propositional knowledge of a belief involving a demonstrative (‘this’) that refers to one’s state of mind, and in this case one’s justification for the belief might also depend on one’s mental state; this is clearly very difficult to communicate to others. Or, if we go for the knowledge-how interpretation, the exercise of the ability to imagine a mental state is, unlike the exercise of the ability to ride a bicycle, not witnessed by others, so one cannot easily demonstrate the possession of such an ability. This is why on both of these interpretations, one might be unable to satisfyingly prove the possession of experiential knowledge to others.

We are left with propositional knowledge and knowledge-how. I cannot see why we shouldn’t say that experiential knowledge involves both propositional knowledge (about an experience) and knowledge-how (to imagine and recognise the experience), in which case we do not have to adjudicate between the two options. Still, even if this does not matter for the purpose of my argument, I think that the knowledge-how has priority, that is, that experiential knowledge is primarily a form of knowledge-how. The reason for this is that, in order to have a belief about an experience, one needs to be able to imagine that experience. (This is not to say that every time one brings to mind the belief, one has to imagine the relevant experience.) Indeed, let’s take the proposition that ‘grief feels very much like fear', as the famous first sentence of C. S. Lewis' *A grief observed* roughly claims (Lewis 2015 [1961]). Assuming that this sentence is true, what does it take to *know* it? Well, presumably, there has to be some way of checking, but if one is not able to imagine the two emotions (i.e. grief and fear), there is a sense in which one does not fully know what 'grief' and 'fear' mean, or does not have the same grip on them as someone who can imagine the two emotions. Therefore, the ability to imagine is essential for having the propositional knowledge.

To conclude, experiential knowledge of an emotion consists in the ability to imagine (correctly) and recognise that emotion, to which we might add some propositional knowledge about how that emotion feels. It follows that there is no *a priori* reason why having had the experience gives better knowledge of what it's like. Just having imagined it, when reading a work of literature, might be just as good if it leads to a good knowledge-how regarding that emotion.

There is though an intuition that if one has experienced an emotion, one's knowledge is better than that of another person who has just imagined it. Kajtár puts this point regarding Mary who reads *The Road*:

> Knowledge does not only admit of types, it also admits of degrees, at least for certain types, and ordinary language allows for this. Certainly, Mary does not know *as much* about fearing death as someone who actually experienced his or her life in danger, but she knows *more than* she did before reading *The Road*. (Kajtár 2015, p. 342)

I share Kajtár's intuition that just by reading a novel and imagining an emotion we usually don't gain as good a knowledge as by having the experience in real life, yet I disagree with him that this is a conceptual matter. In short, I do not think it is always the case that imagining an emotion gives worse knowledge than having that emotion. Rather, I think this is at best a contingent fact that we usually gain better knowledge by having the emotion. To investigate this, we should start by asking, what would make experiential knowledge of an emotion E better or worse? One criterion could plausibly be about how close to E one can imagine an emotion. If one can imagine an emotion that is in the same ballpark as E, yet not very close to it, one has only partial knowledge of what E is like. Another criterion relates to the fact, argued in chapter 3, that imagining an emotion can be schematic, with some phenomenal aspects of the emotions left out. For instance, in fear of a bear that one sees, its claws might appear particularly threatening, with all of the bear's power concentrated in them, ready to hit one; in imagining that emotion, the imaginer might leave out some of these phenomenal aspects and in this way imagine the emotion schematically. Now, it is plausible that for an emotion E, involving some phenomenal idiosyncrasies, the more one is able to imagine these, the better knowledge one has of what emotion E is like. If one imagines E very schematically, one has only partial knowledge of what E is like. Returning to our discussion, there is no reason why someone who has had an emotion E is necessarily able to imagine it better, with more details, than someone who has just imagined it. Perhaps the person who has just imagined it when reading a book might have lingered on it and got a better grasp of the details, and thus be able to imagine it better than someone who has actually had it. It could be that in the vast majority of cases, people who have had the emotion know better what it is like than people who have just imagined it, but this is, importantly, a contingent fact and not, as Kajtár claimed, a conceptual one. Reading a novel and imagining the emotions of its characters can, in principle, give as good a knowledge of what those emotions are like as anything else.

6. THE LIMITS OF THE THEORY

Having argued that literature can give us experiential knowledge by imagining emotions, we can now ask how important this knowledge is and to what extent the theory accounts for what literature can teach us.

As argued in the first chapter, knowing what emotions a person has does not amount to understanding that person. This is because a person can feel the same emotion (or at least very similar emotions) as a manifestation of very different attachments or other drives. For instance, one's hope that Jane become a good lawyer might be a manifestation of one's attachment to Jane; or of one's attachment to one's school, which has a tradition of producing good lawyers and of which Jane is also an alumna; or of one's commitment to the practice of law, Jane seeming very promising in this sense. The agent herself, even after having many emotions, might not be able to reconstruct what attachments or other drives her emotions are manifestations of, and another person who just observes the agent is in this sense in an even worse predicament. Therefore, knowing what the hope that Jane becomes a good lawyer is like does not amount to understanding the person experiencing that hope.

Similarly, having a good grip on a character's fear of death does not amount to understanding what drives that character, for people have very different drives that lead to them wanting to stay alive – attachments to other people, to one's mission on Earth that hasn't yet been fulfilled, to the cultural milieu one inhabits, or just to worldly pleasures. It is not a great surprise that Kajtár chose, to exemplify his view, McCarthy's *The Road*. This novel avoids particularisation and tries to depict what I take to be a common experience, not in the sense of an experience that many people actually have, but in the sense of an experience that many people *would* have if facing the circumstances of the novel. It is clear that the father and the son are attached to each other, but there aren't many details of their attachment, so that the experiences and dialogues depicted could be of almost any father and son that are attached to each other. Any other drives that the father might have, any other things that he might have cared about in the world before the disaster and that he now wistfully remembers are barely hinted at. The point of view of the novel might be taken to be that it doesn't quite matter what drives the emotions of the two characters, as many different sets of drives might have led to the same emotions.

But of course, many novels are not like this one; besides giving us an idea of what the emotions of the characters are like, they try to make us understand the characters, that is, to understand what it is that motivates them and to see their emotions as stemming from the respective drives. This is not so much a value judgment, to the effect that George Eliot or Marcel

Proust are better than Cormac McCarthy, as an observation that many novels, in particular the 19th or early 20th century ‘realist’ novels, have this other project. It is this sort of project that I will account for in the next chapter.

7. CONCLUSION

I have argued that literature gives us experiential knowledge by helping us imagine the emotions of literary characters, in this way allowing us to grasp the phenomenology of many more emotions than we can actually experience ourselves. I have also argued that experiential knowledge of an emotion is primarily an ability to imagine and recognise that emotion. It follows that the knowledge we get from imagining emotions when reading literature is not necessarily worse than the knowledge we might get from having the emotions ourselves. This experiential knowledge does not, however, account for the project of understanding characters that many realist novels have, a project that I will investigate in the next chapter.

CHAPTER 5. LITERATURE AND ATTACHMENTS

1. INTRODUCTION

In this chapter and the next, I will put forward my positive proposal as to what we can learn from literature, or in other words, my own theory of literary cognitivism. In this chapter, I will show how literature, novels in particular, can acquaint us with attachments amongst characters and put us in a position to judge them ethically. I will also argue that this is the best way of understanding the ethical ambition of many works.

However, this does not amount to showing that we learn something *significant* from novel. Judging one instance of an attachment might give the reader knowledge of just one thing, that instance, with no further applications. It is in the next and last chapter where I explain why this knowledge is more important than this.

The plan of the chapter is as follows: I will start with a longer discussion of what I aim to do in this chapter, to give a sense of what we might expect from my theory. I will then go on to put forward my view and illustrate it by discussing Edith Wharton's novel *The Age of Innocence.* I will end by comparing my view with Martha Nussbaum's and arguing that mine better captures the ethical ambition of many works of literature.

2. AIMS

The idea of what we call 'literary cognitivism' is that we can learn something from literature. This is part of the broader question of 'aesthetic cognitivism', or the question of whether we can learn something from art in general, but the theory that I will put forward applies primarily to literature, and perhaps to a lesser degree to other narrative arts such as film and theatre. Now, there is a separate question whether the cognitive value of a work of art, if any, contributes or is related to the aesthetic or artistic value of that work, but I will not attempt to answer this question in the thesis. I only aim to give a theory of literary cognitivism and shall avoid discussions of aesthetic implications of the view that I will put forward.

Of course, I won't try to account for everything that we can learn from literature, nor do I think this would be a wise thing to attempt. There are other theories of literary cognitivism that do not compete in any way with mine. Yet views can also compete with each other, for instance in offering incompatible interpretations of works, and I think in my case Martha Nussbaum's theory is a competitor. In the fourth section I will compare my view with Nussbaum's, with the hope of both illuminating my view further and arguing that it works better than hers in the case of many works of literature.

Methodologically, even if nowadays literary cognitivism is fairly popular[48], I think that we shouldn't start from the assumption that it is the default option and that 'we sort-of know that we learn something from literature and just have to explain what'. This assumption might tempt us to argue for something vague like 'literature can teach something about death'. Instead, I think we should take as the default option that, unless we manage to account clearly and convincingly for what we might learn from literature, we don't learn much, if anything at all.

In building my view, I aim to show not only that we can learn something from literature, but also that this form of learning is important in two senses. First, it is important in the sense that it is hard to learn the same thing in other ways. In particular, I want to show how the details of a work are important and how a sketch of the work would not afford the same insight. Second, this learning is important in that the knowledge we gain is important ethical knowledge and not a mere negligible extra that does not change much in our lives.

Lastly, I think that even though I will argue that many works of literature can give us ethical knowledge, this doesn't mean that these works indicate that something is clearly good or something is clearly bad. Indeed,

[48] Robinson (2005), Gibson (2007) and Mikkonen (2021) are some of the book-length cognitivist defences from this century.

as Christopher Hamilton argues, ‘it is not always clear just what the moral character of certain works of art is, just what the moral view or vision is into which they invite one, or which they express or evince’ (Hamilton 2003, p. 47; cf. Kundera 1988). This might be true even of works about which we feel on every page that they are ethically charged. Indeed, as we shall see, the insight that some works afford us cannot be captured in a thumbs-up or thumbs-down verdict.

3. A THEORY OF LITERARY COGNITIVISM

In this section, I shall put forward my view, present it with short examples here and there and then exemplify it at large by discussing how it applies to Edith Wharton’s *The Age of Innocence*. Essentially, the thesis that I am putting forward is that literature, novels in particular, can acquaint us with attachments of characters and put us in a position to understand their attachments and to judge them.

First, we should recall from the first chapter that an attachment is a drive directed at the person (or thing) one is attached to, where by ‘drive’ I mean a standing mental state that encapsulates what the agent cares about and which manifests itself in emotions. I argued that the drive is distinct from the emotional disposition it manifests itself in, but, given that the drive is an unconscious mental state, we get to understand it by understanding the emotions it manifests itself in and piecing them together.

Now, a work of literature can draw us into imagining the emotions of characters, an imagining which, as argued in the third chapter, consists in forming thick meta-representations of those emotions. It can also draw attention to some relevant emotions and encourage us to dwell on them. Lastly, the episodes it depicts can be revealing in such a way that we can piece together the emotions that we have imagined into an understanding of the attachment they are a manifestation of. This would mean that the novel acquaints us with the attachment.

Of course, there is a sense in which we do not have full acquaintance with the attachment. If we were acquainted with further emotions, we would slightly change our general picture of the attachment. But the idea is that a work of literature can give us a good enough acquaintance, such that we are in a position to make a judgment. Moreover, and importantly, I take it that at least in many works there is an implicit convention that the acquaintance is not *deceptive*, in the sense that the emotions that we are encouraged to imagine do not push us towards a mistaken impression of the attachment. To understand what I mean by this, I will discuss, as an example, Jane Austen's *Emma* (Austen 2015 [1816]). The eponymous heroine, Emma Woodhouse, forms a romantic attachment to family friend George Knightley, an attachment that might have evolved, it is suggested, from a friendship-like attachment. Because of her haughtiness and tendency to lord it over the world around her, it is only towards the end of the novel that Emma realises she has this attachment. The novel acquaints us with this attachment, yet, as just mentioned, it is not a full acquaintance. Now, let's imagine two additional scenarios – that do not appear in the novel – and think of the way they would change our minds about this attachment.

First, consider the imagined scenario in which Emma teased Knightley that, unlike her father, he doesn't play backgammon, so she cannot amuse him by playing backgammon with him. Such a scene does not appear in Austen's novel, but if it had appeared, it would have pushed us to imagine Emma amusement at Knightley's charming seriousness and that would have added a new dimension to her attachment to him. However, the scene is also consistent with what happens in the novel: Emma plays backgammon with her father and we aren't told anything as to whether Knightley plays backgammon. Knightley is the kind of character who could find backgammon a bit silly and Emma is the kind of character who would tease people. More to the point, Emma's attachment to Knightley, unlike that towards her father, is of a kind that might lead her to tease him. So, such a scenario, depending how it is expanded upon, would change our understanding of Emma's attachment to Knightley a bit, but by *adding a new aspect*, not by forcing us to reinterpret the acquaintance we already have.

However, there could be potential scenarios, again not hinted at in the book, but strictly speaking compatible with what is written, that would change our understanding of what drives Emma. One such scenario would involve Emma demanding Knightley that he renounce all sorts of other endeavours of his – for instance, the management of his lands – to concentrate solely on her, a scenario which would lead us to think of her attachment as more possessive than we thought. We might then reinterpret her happiness at the sacrifices Knightley made for her – for instance, that of moving together with her and her father – not as sheer happiness at the fact that her beloved loves her back to such an extent, but rather as happiness at furthering her control over him. It's not so much that we would imagine her happiness differently (though arguably we might imagine it slightly differently), but the important thing is that we would take the drive that it stems from to be different and so understand the happiness differently. Such a scenario would not merely add another aspect to our understanding of Emma's attachment, but would force us to reinterpret what we know.

Another such scenario, which would disrupt the understanding got from the book, would involve Emma preventing her father from spending time with other people. For instance, suppose that her father found someone to occasionally play backgammon with and that Emma did everything she could to ensure that her father would play only with her and not with his other partner. Such a scenario would make us conjecture that Emma has a certain general possessiveness as a drive and, furthermore, make us wonder to what extent this drive was effective in her decision to marry Knightley. If we assume that it was effective, this might lead us to conclude that her attachment to Knightley was weaker than we thought and that some of her behaviour and emotions towards him are partly accounted for by this possessiveness. Just to clarify, in this scenario, the possessiveness would be a separate drive that would make us conjecture that her attachment to Knightley is weaker than we had thought; in the previous scenario, we would conjecture that her attachment, though as strong as we thought, was in other ways different from what we had gathered from the book. The last two scenarios, unlike the first one, in which Emma teases Knightley for not liking backgammon, would come across as *surprising*; the reason for this is that they contradict, or at least somewhat undermine, the picture of what drives the characters that we formed from reading the novel.

Where I am trying to get with this long-winded discussion is that I think there is an implicit convention in much literature, in particular in realist novels, that the attachments of characters are such that episodes of the first kind, in which we would understand a new aspect, could happen, while episodes of the second kind, in which we would reinterpret what we've read, could not happen. Note that I am saying 'could(n't) happen' and not 'do(n't) happen', hence I am not claiming that we need to take a stance as to what happens in the fictional world outside what is depicted in the book – I remain neutral on the question of what, if anything, is true in the fictional world besides what is depicted. Also, again just to clarify, I am not saying that such scenarios could not appear in a realist novel – there are realist novels which at various points force us to reinterpret what we've read so far. Instead, the claim is that *if* they do not appear in a such novel, we should assume that the drives of the characters are such as scenarios like these could not happen. To pre-empt a possible worry, I think the question of which of the scenarios described could happen is not absurd or philistine, it is not of the 'how many third cousins does Emma have?' sort. This sort of question relates to our understanding of the characters as depicted in the novel and to our judgments of them, and it is for this reason that we have to consider these potential scenarios. Summing up this discussion, the picture that we form of characters' attachment from the novel might be incomplete, but it is not misleading; therefore, if it is revealing enough, it can give us a good idea and allow us to pass a judgment.

A similar discussion can be had about the emotions that we imagine characters to have. In the third chapter, I fave argued that we can imagine some emotions schematically, that is, without all the phenomenological details. This can apply to our imagining fictional characters' emotions and it might sometimes be encouraged by the work itself, as works dwell on some emotions, allocating them rich descriptions, while mentioning some other emotions in passing. Following a similar reasoning as before, I take it that if the work mentions an emotion in passing and does not encourage us to imagine its phenomenology in detail, there is an implicit assumption that a less schematic imagination of that emotion would perhaps add a new aspect, but would not force us to reinterpret the picture we already have of the character and her attachments or other drives.

Now, the works of literature that I am focusing on, which are mostly novels, have a narrative form. They are formed of episodes that succeed each other and are in some way connected to one another. I now want to show why the narrative form is important by highlighting two ways in which it can be so.

First, some of the characters' emotions might be hard to imagine without a significant background knowledge of what has happened before in the work. To imagine these emotions well, the reader needs to have read until the relevant passage and use her knowledge of previous episodes. To see what I mean by this, let's take an example from Joseph Conrad's *The Secret Agent*. In the novel, we find out that Winnie Verloc has married Mr Verloc mostly because she thought he is a good enough man that could care for her and, *most importantly*, also for her brother Stevie, who is portrayed as a 'simpleton', who could not take care of himself. However, Mrs Verloc's impression is completely shattered when Stevie dies absurdly, because of Mr Verloc, with the latter not seeming to care about this. After she finds this out, the following passage occurs:

> A few seconds only had elapsed since the last word has been uttered aloud in the kitchen, and Mrs Verloc was staring already at the vision of an episode not more than a fortnight old. With eyes whose pupils were extremely dilated she stared at the vision of her husband and poor Stevie walking up Brett Street side by side away from the shop. It was a last scene of an existence created by Mrs Verloc's genius; an existence foreign to all grace and charm, without beauty and almost without decency, but admirable in the continuity of feeling and tenacity of purpose. And this last vision has such plastic relief, such nearness of form, such a fidelity of suggestive detail, that it wrung from Mrs Verloc an anguished and faint murmur, reproducing the supreme illusion of her life, an appalled murmur that died out on her blanched lips.

> 'Might have been father and son.' (Conrad 2012 [1907], ch. 11, pp. 211-212)

To understand this passage, to imagine what Winnie Verloc feels, we need to have read the novel until then and have in the background of our mind the scene in which Mr Verloc leaves with Stevie, as well as other scenes of them together, an image of the shop which the Verlocs ran and so on. Such a passage can only come late in the novel, precisely because in order to imagine what Mrs Verloc feels (and for the cruel irony to have its effect, but this is for another time), we need to have quite a bit of background. Of course, this passage and what follows from it serve to deepen our understanding of Winnie Verloc's attachment to Stevie; but in order to imagine the emotion that leads to this deepening of our understanding, we have to already have a grip on what significance many things have for her.

The second reason why the narrative form of literature might be essential for our understanding the attachments of characters is that, as argued in the first chapter, once an attachment is formed, it contains the potential for its own evolution. If X has an attachment to Y, how that attachment changes in virtue of X finding out something unexpected about Y, or in virtue of events they go through together, is influenced by the attachment itself. Had X found out the same things about Z, to whom she has a different kind of attachment, her attachment to Z would have changed in a different way. It follows that in order to understand an attachment, we need to understand how it evolved and a work of literature can depict its evolution.

To make everything that I've discussed so far clear and plausible, I will discuss in detail one novel and what understanding my theory claims it provides. The novel I chose is Edith Wharton's *The age of innocence* (1999 [1920]) and the reason for this choice is threefold: first, it is a very good novel that exemplifies the points made above; second, it has been already discussed in the philosophical literature, notably by D.Z. Phillips (1972); and third, it depicts some attachments that, as Phillips observes, might strike some readers at first sight as strange or even incomprehensible, and, while I am not claiming that we only learn something from attachments that seem strange to us, this serves to make my point more easily.

The plot of the novel is as follows: Newland Archer is a member of the 'old New York' elite, a mid-late 19th century society ruled by a strong sense of decorum, hierarchy and tradition, that is initially ironically described by Wharton as 'a kind of hieroglyphic world, where the real thing was never said or done or even thought, but only represented by a set of arbitrary signs' (ch. 6, p. 29). He is engaged to May Welland, with whom he is in love 'sincerely but placidly' (ch. 6, p. 29). Overall, he is attached to both his society and to May, but occasionally feels that they all lack in sophistication and imaginativeness – for instance, in their lack of artistic preoccupations, at least compared to how things were in Europe at the time:

> he had often pictured to himself what it would have been to live in the intimacy of drawing rooms dominated by the talk of Mérimée (whose 'Lettres à une inconnue' was one of his inseperables), of Thackeray, Browning or William Morris. But such things were inconceivable in New York. (ch. 12, p. 66)

This equilibrium is shattered by the arrival of Ellen Olenska, another member of the New York society who has lived abroad for a long time, a cousin of May, and now the estranged wife of a Polish count. Newland helps her be accepted again in society and advises her against divorcing her husband, in order to protect both her reputation and that of her family. Initially, he seems to do this because Ellen and May are cousins, but then he starts sympathising with her. She seems to embody all the sophistication and imaginativeness that is lacking in New York: Newland talks to her about all exhibitions in London and Paris and observes that she has books in her drawing room – that is, not just in her library, as everyone else. They end up falling in love. Thinking of eloping together, they conclude that this would be a betrayal of their families, of all the people they care about and of their society, and they decide to sacrifice their relation. Ellen goes back to Europe, while Newland marries May Welland.

I take it that one of the aims of the book is to reveal what it is that motivates Ellen and Newland. As D.Z. Phillips (1972) argues, many a modern reader might see their decision not to elope as a form of weakness, of being bound by an exterior force, by a taboo, by what society tells them, etc., a construal determined by the assumption that, at least in matters of the heart, either one is striving for happiness or one acts in virtue of a taboo that prevents that striving. Instead, Phillips argues that we should 'wait on the novel' and discover what actually motivates the characters (1972, p. 55). His claim is that they are not motivated by taboos, external pressures and so on, but rather they just have different values, 'values involving suffering, denial, endurance, discipline' (1972, p. 58). I agree with this in principle, but as it stands it is a bit vague – what does Phillips mean by 'values'? He sometimes talk about 'Newland Archer, Ellen Olenska and May Welland, in different ways, embody[ing] the old New York morality Edith Wharton wanted to depict' (1972, p. 58). If embodying a 'morality' means believing that there are some set of rules or principles that *persons in general* should abide by, then it is not clear that this is the case. And it does not seem that Phillips thinks in this way either. Surely, we can say that they embody the old New York values in the sense that their lives might have been praised in that society, but if we want to understand them, we need to understand what drives their lives and the decisions they make.

In short, what I want to claim is that Ellen and Newland are motivated in what they do by two types of drives. The first type consists in their attachments to the society they are part of, the 'old New York'. The second type consists in their attachments to each other, attachments that, as we shall see, play a more subtle role than we might have initially thought. By reading the book, we are supposed to understand these drives and what role each of them plays. Let's take each of them in turn.

First, the attachments to their society. Newland is described as always having had a somewhat sacrificial attachment to his society, which led him to embrace its practices and contribute to its structure. He does think that it is somewhat unimaginative and that some members manage to get away with caddish behaviour under the cover of form, but overall, he is committed to his society and is ready to take risks for its well-functioning –

in short, he accepts his mother's idea that 'if we don't all stand together, there'll be no such thing as Society left' (ch. 6, p. 33).

Ellen, on the other hand, has at the beginning of the book a more carefree attachment, which led to her just being happy to come home, to be amongst her people, after so many years of being away. She is hardly aware that some people see her with unfriendly eyes for being estranged from her husband – for instance, by frowning upon the fact that her family allows her to sit in their box at the opera. Even regarding Newland, who is sympathetic to the person whom everyone refers to as 'poor Ellen Olenska', we are told that

> he was glad that his future wife should not be restrained by false prudery from being kind (in private) to her unhappy cousin; but to receive Countess Olenska in the family circle was a different thing from producing her in public, at the Opera of all places, and in the very box with the young girl whose engagement to him, Newland Archer, was to be announce within a few weeks. (ch. 2, p. 8)

Ellen's family back her though and through various schemes convince the van Luyden family to invite her for dinner. The van Luydens are one of the most prestigious families, whose stamp of approval means a lot in the old New York, so the invitation for dinner seals her definitive acceptance back in the society, an acceptance that everyone has to acquiesce to. I take it that all these efforts and sacrifices that are done for her serve to change her attachment to the society into one that is more prone to sacrifices. Note, though, that, as I have argued in chapter 1, the change of attachment is itself a manifestation of the attachment, which contains in itself its own potential for development. So, the original attachment of Ellen was decisive for its own evolution.

Even though Newland was the main person who contributed to how Ellen changed and who originally upheld his attachment to the New York society, it is he who proposes that they elope. Ellen remonstrates that that would be wicked and a betrayal of what they both care about and, like a recent convert, she explains to him:

> ‘Just imagine,’ she said, ‘how stupid and unobservant I was! I knew nothing of all this till Granny blurted it out one day. New York simply meant peace and freedom to me: it was coming home. And I was so happy at being among my own people that everyone I met seemed kind and good, and glad to see me. But from the very beginning,’ she continued, ‘I felt that there was no one as kind as you; no one who gave me reasons that I understood for doing what at first seemed so hard and — unnecessary. The very good people didn’t convince me; I felt they’d never be tempted. But you knew; you understood; you had felt the world outside tugging at one with all its golden hands – and yet you hated the things it asks of one; you hated happiness bought by disloyalty and cruelty and indifference. That was what I’d never known before – and it’s better than anything I’ve known.’ (ch. 18, p. 110)

Seeing her so steadfast in her commitment to their society consolidates Newland’s attachment to it as well. They give each other up, and Ellen returns to ‘Europe’ (by which they seem to mean France).

Importantly, I think it would be incomplete to say that Ellen and Newland were motivated by their attachments to their society alone. Another very powerful driving force is their attachments to each other. It is an important part of their love for each other that it was formed in this society and, so to speak, under its auspices. The fact that they are both attached to it was integrated in their attachments to each other as a kind of foundation – an important thing that bound them together. It is therefore their very love for each other that brings in the tragic tension – eloping would be a betrayal of one of its foundations, as Ellen’s remonstration shows:

> ‘Ah, don’t let us undo what you’ve done!’ she cried. ‘I can’t go back now to that other way of thinking. I can’t love you unless I give you up.’ (ch. 18, p. 110)

‘I can’t love you unless I give you up’ is the key line that synthesises their predicament. Their attachments push them to elope, but they seem to push more strongly in the other direction, impelling them to part, so their decision to renounce each other is not solely the manifestation of their attachments to their society.

It is in this light that we should see the end of the book. Many years after the events just described, after his wife May passes away, Newland is in Paris with his son Dallas and by chance gets an opportunity to see Ellen again. Wandering around Paris, he muses about what might come:

> A few streets away, a few hours away, Ellen Olenska waited. She had never gone back to her husband, and when he had died, some years before, she had made no change in her way of living. There was nothing now to keep her and Archer apart – and that afternoon he was to see her.
>
> […]
>
> ‘But I’m only fifty-seven – ’ and then he turned away. For such summer dreams it was too late; but surely not for a quiet harvest of friendship, of comradeship, in the blessed hush of her nearness. (ch. 34, p. 226)

The father and son go to her apartment building, but when about to enter, Newland tells his son that he will wait a bit longer outside. The scene unfolds:

> Dallas looked at him again, and then, with an incredulous gesture, passed out of sight under the vaulted doorway.
>
> Archer sat down on the bench and continued to gaze at the awninged balcony. He calculated the time it would take his son to be carried up in the lift on the fifth floor, to ring the bell,

> and be admitted to the hall, and then ushered into the drawing-room. He pictured Dallas entering that room with his quick assured step and high delightful smile, and wondered if the people were right who said that his boy 'took after him'.
>
> Then he tried to see the persons already in the room – for probably at that sociable hour there would be more than one – and among them a dark lady, pale and dark, who would look up quickly, half rise, and hold out a long thin hand with three rings on it… He thought she would be sitting in a sofa-corner near the fire, with azaleas banked behind her on a table.
>
> 'It's more real to me here than if I went up,' he suddenly heard himself say; and the fear lest that last shadow of reality should lose its edge kept him rooted to his seat as the minutes succeeded each other.
>
> He sat for a long time on the bench in the thickening dust, his eyes never shone through the windows, and a moment later a man-servant came out on the balcony, drew up the awnings, and closed the shutters.
>
> At that, as if it had been the signal he waited for, Newland Archer got up slowly and walked back alone to his hotel. (ch. 34, pp. 228-229)

The final scene is a key one, that rounds off the novel in a remarkable way. Having reached that point in the story, we already have a good understanding of Ellen and Newland's attachments to each other. But now we understand something more, something that perhaps not even Newland understood before this scene: the sacrifice they made actually bound them together even more, yet their attachments evolved to the effect that anything they might do together now, after many years, would be a kind of sham. When visualising the scene inside the building, Newland feels her presence, yet it is her presence as she was many years ago, and seeing her for real would eliminate this presence and 'that last shadow of reality should lose its edge'. In imagining the emotions that Newland feels in that moment, we get to understand, together with him, how his attachment to Ellen has evolved.

Newland's decision not to go up is not, as one might superficially think, an expression of his attachment to Ellen being weak but of its being strong. As argued in chapter 1, the strength of an attachment should not be cashed out in terms of, say, how much time one is willing to spend with the other person, but in terms of how much motivational power the attachment has. The actions that it motivates need not be towards spending time with the other person and can also pull in the other direction. It is like this for Newland, whose strong attachment prevents him from pursuing a pleasant but ultimately superficial encounter.

In order to imagine as we should what Newland feels when sitting in front of the building, we need to have read the novel. Everything he visualises, he visualises in terms of his history with Ellen, in terms of how her New York drawing room was filled with her presence when Newland saw her there. If we hadn't read the previous scenes in which she sits on a sofa corner, with flowers behind her, we struggle to imagine the relevant details of Newland's emotion. Note that in describing what goes on in Newland's mind, Wharton does not even use many terms that denote emotions, except in the paragraph about its being 'more real to me here'. Just the focus on the things Newland visualises is enough to convey to the reader who has followed the story until then what he feels.

Note also that, as mentioned in the second chapter, an emotion can indicate two incompatible courses of action, and sometimes it can be hard to understand, from a third-person perspective, how the courses of action stem from the same emotion. Here we have a very good example of this: what Newland feels when sitting in front of Ellen's building pushes him to go up, but also to stay put, with the latter push being stronger. By imagining his emotion in that moment, the two courses of action appear intelligible as stemming from the same emotion.

We can also see why the temporal dimension of the novel is essential to understanding the attachment that Newland has to Ellen at the moment of the last scene. As mentioned before and argued in the first chapter, one's attachments change in reaction to new events, discovered facts and so on, but the changes are not imposed on the attachments, as it were, from the outside. Instead, the changes are driven by the attachment itself. So in order to understand Newland's attachment at the end, we need to understand how it was before the decision that he made with Ellen to sacrifice their relationship and how it evolved in reaction to that decision. But in order to understand how it was before they made the decision and how it lead them to make that decision, we need to understand how it was formed and what role their attachments to their society played in that. If someone who hasn't read *The age of innocence* cannot understand how a man, without any constraints, can refuse to go up and meet the woman he has a strong attachment to, it would help rather little to try to describe what emotions Newland had at that moment. The person would find it hard to imagine them, and if they do, they would find all this rather puzzling. Instead, what we should do is explain them how Newland *ended up* having this attachment by describing the evolution of the attachment – in short, we should give them the book to read.

It would now be useful to compare *The Age of Innocence* with two other novels that are superficially similar, not in terms of style, but in terms of the broad shape of the story: Madame de Lafayette's *The Princess of Clèves* (2020 [1678]) and Evelyn Waugh's *Brideshead Revisited* (1978 [1945]).

In *The Princess of Clèves*, the eponymous heroine is married to a good man whom she admires but does not quite love (or anyway, whom she loves more as a friend). She falls in love with the duke of Nemours, yet refuses to become his mistress and tries to avoid him – *even after the death of her husband*. Her predicament and her choices are strikingly similar to those of Newland, so someone might lazily conclude that the two novels 'portray the same values, even though they were written almost 300 years apart'. I believe that this would be a mistake: what motivates the princess' actions and shapes her life is a kind of ideal of virtue, one that would include chastity, a distrust of one's instincts and perhaps a form of self-sufficiency. The novel does not seem to suggest that her attachment to the 16th century French court or even to the duke of Nemours played a part in her key decisions. Rather, all her attachments were less strong than her aspiration for a certain kind of virtue.

A similar story can be told about Waugh's *Brideshead Revisited* (1978) [1945]. After being married to the wrong people, Charles Ryder and Julia Flyte finally get a chance to be together. Yet it is at that moment that Julia regains her Catholic faith, which, the novel seems to suggest, has been dormant in her and survived all her attempts to get rid of it, and decides that she has to renounce what she loves most. She tells Charles about

> the bad thing I was on the point of doing, that I'm not quite bad enough to do; to set up a rival good to God's. Why should I be allowed to understand that, and not you, Charles? It may be because of mummy, nanny, Cordelia, Sebastian – perhaps Bridey and Mrs. Musspratt – keeping my name in their prayers; or it may be a private bargain between me and God, that if I give up this one thing I want so much, however bad I am, he won't quite despair of me in the end. (bk. III, ch. 5, p. 373)

Of course, I am not going to try to explain what faith is in a paragraph, and the whole novel is a highly complex attempt to understand something about it, but it is clear that we're talking about, using my language, radically different drives from those involved in *The age of innocence.* Any superficial similarity in the plot is just that: a superficial similarity. What the characters care about is completely different in the two novels.

I hope these comparisons help show that the theory I have built in the first half of the thesis serves us well to differentiate these novels and understand that what drives the main characters is very different in all three of them. Even if the decisions might be similar and even if some of the emotions involved might be to some extent similar, it is what ultimately drives the characters that matters and that we should try to understand and, to some extent, judge.

In light of this, Phillips' original observations show their limitations. Indeed, we might say of all the three novels that the characters are driven by 'values involving suffering, denial, endurance, discipline' (Phillips 1972, p. 58), but this would give the wrong impression that the drives involved are somewhat similar and obfuscate the major differences between the three novels.

Now, I take it that some, though perhaps not all, novels also pull us, more or less subtly, towards an ethical judgment of the attachments they depict[49]. *The Age of Innocence* makes no exception, it seems to me, as it is ultimately sympathetic to the main characters and admires their attachments, not including any point of view that challenges them. It is not sympathetic unreservedly though and does, at various points, portray the society the two are attached to as somewhat claustrophobic. It is no surprise, therefore, that people read the novel in very different ways: the motivation for awarding the Pulitzer prize to the novel says that it depicts 'the wholesome atmosphere of American life and the highest standard of American manners and manhood', while a recent laudatory review says: '*The Age of Innocence* makes an ironic commentary on the cruelties and hypocrisies of Manhattan society in the years before, during and after the Great War' (McCrum 2014; Pulitzer quote is taken from there)[50].

We can now come back to an observation made in the previous section. A work like *The Age of Innocence* seems not to try to convince us of some general ethical proposition, but to acquaint us with some particular lives, that involve particular attachments, with all their idiosyncrasies, and encourage us to evaluate these attachments. The work itself might pull us towards certain judgments, but this pull is not always obvious and may often be very subtle. Moreover, there might be a pull towards opposite judgments as well, resulting in the kind of ambivalence mentioned in the previous section, that it is sometimes hard to tell, even about an ethically charged work, what exactly is the ethical view it proposes (Hamilton 2003). At the risk of making too sweeping and vague a statement, I would say that in many such novels it feels that the writer wants to talk to us about something that they feel it's important to understand, think about and evaluate, but

[49] This is a huge topic though, and I do not want to enter into debates as to how works push us towards ethical judgments. The idea that works of literature prescribe or at least encourage ethical judgments is important in discussions of what the ethical (not cognitive) value of works of art is and how it relates to their aesthetic or artistic value – see Gaut 1998, Carroll 1996, Kieran 2003a.

[50] Critical interpretations also tend to be divided along these lines: Auchinsloss (1962) is an example of the first kind, and Stuart Hutchinson's introduction to Wharton 1999 [1920] of the second.

without indicating a very clear judgment. I have argued that one way to cash out this ‘something’ is as attachments of characters, that we come to understand and evaluate in the works.

4. COMPARISON WITH MARTHA NUSSBAUM’S VIEW

I will now compare my view with Martha Nussbaum’s (1990) view of what we can learn from literature, trying to show that mine accommodates better the ethical ambitions of many works, including those discussed by her as well as by me. She exemplifies her view with another novel, Henry James’ *The Golden Bowl*, but I take it that her analysis could be readily applied to *The Age of Innocence* as well, and mine to *The Golden Bowl*, so we can compare the two views.

Nussbaum starts from a complex Aristotelian picture. According to this picture, very often, when faced with a decision, an agent cannot just reduce the two options to a common measure (such as utility), because the goods involved are incommensurable. A common example is a conflict of demands between helping a friend and a performing a public duty. The virtuous agent should recognise the incommensurability, yet she also has to make a decision. So, how does the virtuous agent make the decision? Nussbaum concedes that there are some moral rules that often hold, but claims that these rules are more like generalisations from particular situations and that the agent should judge each situation individually. In judging them, the agent should use what Nussbaum calls ‘perception’, by which she means a ‘complex responsiveness to the salient features of one’s concrete situation’ (1990, p. 55). This responsiveness involves, besides reasoning, emotions that pick up various salient features of the situation and recognise the value involved in those salient features. It also involves imagining various aspects of the situation, such as the perspective of other people involved, seeing the decision in a larger pattern of events, etc. I take it that for Nussbaum all these mental processes are not mere means to virtuous actions, but actually constitutive of virtuous actions. To perform a virtuous action, it is not enough to land (perhaps by chance) on the right decision – one needs to arrive at the right decision in the right way, involving all the elements of perception described above. Similarly, to realise, from a third-person perspective, what the virtuous action in a certain situation is, one should perceive the relevant features of the situation in the right way.

To illustrate this, we might think of the situation that requires intervening in a conflict between two of one's neighbours. One should try to appease the conflict, for the sake of the well-functioning of the community, but at the same time one should give each of the conflicting parties their due, which can be tricky, especially if the conflict is actually an escalation of legitimate complaints on both sides. Also, one should not treat solving their conflict as a mere means to the well-functioning of the community. This is just a sketch of all the problems involved in such a situation, and it is rare that two such situations are really alike, so we cannot completely codify what one should do – one needs the kind of perception invoked by Nussbaum in order to feel when there's a risk of being too intrusive, to imagine the perspectives of the participants, to have an intuitive sense of how to foster the well-functioning of the community or to see how legitimate the complaints are. (It's not for nothing that having neighbours can be quite a challenge.)

Now, there are potentially unlimited aspects of a situation and even more ways to apprehend those aspects or think about them, so virtue is no easy thing. To become virtuous, one must develop one's perception, and this can require a lot of of practice.

Here is where novels come in, at least the good ones, which do exactly this: they help us develop our perceptions by putting characters in difficult situations and showing how they reach the right decisions, how they perceive the situations they face. They draw us into perceiving the situations the characters face as a virtuous Aristotelian would and hence teach us to perceive. Even more, by depicting situations that differ from one another in substantial ways, and showing how the virtuous actions in all these situations cannot be captured by principles, novels further enforce the Aristotelian picture described above. They show that the perception that

Nussbaum talks about is not a mere extra quality, one that makes one slightly more virtuous, but instead is central to virtuous action. In short, novels both show us the complexity of ethical life and help us navigate it.

As mentioned before, Nussbaum uses Henry James' *The Golden Bowl* as her main example. In this novel, Maggie Verver is facing a hard decision: in order to truly become prince Amerigo's wife, in fact and not just on paper, she needs to break away from her father, with whom she has so far lived in a sort of Edenic bliss, in which there were no conflicts, no tragedies, and all decisions were easy. (In short, in living with him, she was still a child.) Yet, of course, she doesn't want to break their relationship, but only to part from him and start her new married life. It is very hard for her to do this, and the process involves successive perceptions of the complex nature of the decision, of her father as a separate person from her, of his pain that would ensue from their parting, of how much he means for her and so on. She is to treat him sensitively, to try to communicate all this to him and to attempt to build a plan together. Importantly, her decision should and does involve a strong sense of regret that she has to part from him and that things cannot go on as before (even if she knows that they *cannot* go on as before, as she has to grow up). All her successive perceptions are richly depicted by James, and we are supposed to learn about all the complexity of her situation from this depiction. Moreover, we are supposed to perceive her situation and its salient aspects in a similar way to her. If this is indeed how an Aristotelian agent should perceive such a situation, then we essentially learn how to perceive it. And in this way, we refine our perception in general, therefore becoming better Aristotelian agents. We are led closer to the Jamesian (and Nussbaumian) ideal of being 'finely aware and richly responsible', a state in which we can ourselves make such subtle distinctions and have such nuanced perceptions.

Now, Maggie's decisions involve a very important attachment, namely, the attachment to her father. What role does this play in the broad picture? How should this impact the perception of the Aristotelian agent? Here is Nussbaum's view on how the Aristotelian agent should construe situations involving love and friendship:

> The particularity of love and friendship seems to demand nonrepeatability in yet a stronger sense [than just seeing a situation as particular]. Good friends will attend to the particular needs and concerns of their friends, benefiting them for the sake of what they are, in and of themselves. Some of this 'themselves' consists of repeatable character traits; but features of shared history and of family relationship that are not even in principle repeatable are allowed to bear serious ethical weight. Here the agent's own historical singularity (and/or the historical singularity of the relationship itself) enter into moral deliberation in a way that could not in principle give rise to a universal principle, since what is ethically important (among other things) is to treat the friend as a unique nonreplaceable being, a being not like anyone else in the world. (Nussbaum 1990, p. 72)

Yet Nussbaum also insists that there is *one* right way for an agent to act in a certain moment, in particular in the case of Maggie in *The Golen Bowl*:

> [I]t is extremely important to insist, once again, that the universalizing we do when we read a novel like this one involves very little generalizing. The person who, reading this scene, concluded from it that 'All daughters should treat their fathers as Maggie treats Adam here,' would have shown herself a blunt reader indeed. The reading I have presented suggests, instead, that 'any daughter with Maggie's history and character who has a father with Adam's history and character (where this would be filled in by a very long and probably open-ended set of descriptions), should, if placed in a situation exactly like this one, respond as Maggie responds here.' It also suggests, more pertinently, 'All daughters should treat their fathers with the same level of sensitivity to the father's concrete character and situation, and to the particularities of their histories, that Maggie displays here.' The universalizing, in the latter case, provided

> not a principle, but a direction of thought and imagination.
> (Nussbaum 1990, pp. 166-7)

So, essentially, in order to be a good agent, one needs to analyse one's situation, a situation that encompasses immediate facts, but also historical facts, including the relationships with people one has attachments to; to perceive the situation in the relevant way, using one's emotions and imagination; and then to come to *the* right decision, a decision that any other ideal agent can recognise as *the right one*. Or, if we assume that there could be a few equally right decisions that an agent could make, again, all ideal agents can recognise these possible decisions as the right ones. This is because the recognition of them as right stems from the virtuous perception of the relevant features of the situation. This seems not to be an extra assumption that Nussbaum makes, but rather it is where her Aristotelian view leads her.

I take it that Nussbaum's view could apply to *The Age of Innocence* as well. We could see Newland and Ellen as Aristotelian agents that struggle to perceive in a virtuous way the difficult, if not downright tragic, situation in which they find themselves, taking into account their common history, their society, how what they do might affect that society and the other people that they care about, what implicit obligations they have accrued and so on. Of course, applying the theory does not mean that Wharton (or someone else) necessarily succeeds in showing the readers how a virtuous agent should perceive various situations and hence teaches them to become more virtuous agents. What it means is that the best way to construe the ethical ambitions of some novels is that they attempt to show us this. Perhaps some fail or succeed only partially.

I will now argue that Nussbaum's view does not succeed in accounting for the ethical ambition of *The Golden Bowl*, *The Age of Innocence* and most of the novels that involve attachments, and explain why my view does better in this sense.

The first problem with Nussbaum's view is her assumption about the role of attachments in the life of the virtuous agent. Essentially, the view she ends up with regarding actions and emotions towards people one has attachments to is a variation of Niko Kolodny's view (2003) that I have argued against in the first chapter. For Nussbaum, as for Kolodny, the defining feature of the attachment is just something exterior to the mind of the agent, namely, the history of the relationship, something that one should *take into account* when making a decision. Let's recall from the quotation above that Nussbaum's view is that

> 'any daughter with Maggie's history and character who has a father with Adam's history and character (where this would be filled in by a very long and probably open-ended set of descriptions), should, if placed in a situation exactly like this one, respond as Maggie responds here.' (Nussbaum 1990, pp. 166-7)

This means that how a virtuous agent should react is determined by the common history and the character of the people involved. But in the first chapter, I have argued that an attachment is a drive, that is, a standing mental state, and that if one has an attachment to another person, one's emotional reactions are a manifestation of that drive and *not* fully determined by external factors. If other people, virtuous or not, are placed in situations like Newland's or Maggie's, they could not act and feel as Newland or Maggie do, for these other people have different drives, including attachments, and how they act and feel is a manifestation of these drives.

Even more, the Aristotelian interpretation of Nussbaum seems to me to attempt to fit James' novel (and Wharton's) into a mould that does not do justice to the novel's ethical ambitions. It is not clear that these novels are best read as depicting *the way* an agent should see the world and act, as putting their characters in situations in order to show, using their responses, how *the ideal Aristotelian agent should respond to such situations.* Rather, these authors seem to depict what Newland Archer and Maggie Verver act and feel, based on their own attachments and their own view of life, and help us understand why *these particular characters*, with their own motivations, act and feel as they do. Of course, the novels implicitly put forward these characters as interesting to understand and judge, perhaps positively, but not as some kind of ideal agents.

My view, on the other hand, acknowledges that these works depict *some ways* to see the world and act. For each character, the way they see the world (in emotions, but also in what they attend to, what they imagine, etc.) is determined by the drives they have, that is, what they care about, and these drives include their attachments. Of course, if the novels have some ethical ambition, the attachments they depict might be in some way instructive, but this instructiveness need not be a matter of the novels putting the attachments forward as *the best* or as *models*. They might even be problematic in various ways, but even these ways can be instructive. For instance, I cannot help but find the relationship between Maggie Verver and her father as having a tinge of *folie à deux*, in which the two think of the world as essentially theirs to do whatever they want with it and reinforce each other's attitude to that effect, in the absence of any feedback from the rest of the world. Yet I see that even in these attachments there is something to be admired.

Where my view fundamentally differs from Nussbaum's is that she seems to ask, perhaps implicitly, what the best way to perceive (and act in) a certain situation. In contrast to this, I have argued in the first chapter that individual emotional reactions and decisions are manifestations of deeper mental states, so the question as Nussbaum puts it is somewhat misleading. In the context of an attachment, what we should primarily judge is not so much a momentary decision, for this cannot be fully understood in itself, but the attachment it is a manifestation of. By introducing the notion of an attachment and showing how emotions are manifestations of it, I have created a framework in which we can easily see why we should not over-concentrate on momentary decisions, asking whether they are right or wrong, but rather focus on the understanding of the characters that only a narrative work can afford and pass some general judgment.

A third worry that I have regarding Nussbaum's view, and where I also think that my theory fares better, is that the view somewhat instrumentalises literature. Despite its subtlety in accounting for why the rich details contribute to the ethical insight that a work might afford, it seems that if the ultimate goal is just to become more virtuous, to improve our Aristotelian perception, then it is important that one reads good novels, but it is not clear whether the particular novels that one reads make too much of a difference. Also, if an agent is virtuous enough, such that they already have a good enough perception and do not need the ‘training’ of the likes of James and Eliot, there seems to be very little that they could learn from these novels. Yet it seems that there is a lot that even such a person could learn from literature.

In contrast to Nussbaum, I maintain that this is not a matter of virtue. There is always something to be learned from being acquainted with and judging a new attachment, an attachment that might be different from those that one has encountered before. Therefore, there doesn’t seem to be a moment when one can say that one has finished learning.

To sum up this section, I think that the view I have proposed in this chapter does much better in capturing the ambition of many novels and that we should not try to fit these into an Aristotelian mould, as Nussbaum does.

5. CONCLUSION

I have argued that literature can acquaint us with attachments that characters have to each other. It does this by helping us to imagine their emotions and to piece these emotions together in order to understand the attachments they stem from. The narrative form is important, as, first, some emotions are hard to imagine without understanding the context in which they appear; and second, and more importantly, the attachments evolve through time based on their own logic, and to understand them as they are at a certain moment in time, we need to understand how they reached that state. Acquainting us with attachments in this way, literature can put us in a position to judge them ethically.

However, it is not clear at this point whether we can be said to learn something important from understanding and judging one or more attachments from a work of literature. Why would this amount to learning something of relevance beyond that work of literature, with its particular characters and situations? In the next and final chapter, I will argue that it can amount to learning something very important.

CHAPTER 6. AN ACQUAINTANCE PRINCIPLE

1. INTRODUCTION

In chapter 5, I have demonstrated how a novel can acquaint us with the attachments of literary characters, showing us how the emotions and behaviour of the characters make sense together and reveal what their attachments are. Now, the final question is, does this mean that the reader can learn something significant from literature in this way? We might agree that laying the ground for such an acquaintance can be a remarkable achievement of the writer and that it can provide the reader with aesthetic as well as intellectual pleasure. But none of this means that there is something important to be learned from the work, that the reader is in any significant way illuminated.

My claim in this final chapter, that this form of acquaintance can lead to important knowledge, is based on a thesis about the relation between the non-evaluative descriptions of attachments and evaluations of them. I will argue for what I will call an 'acquaintance principle', namely, that the value of attachments is revealed in particular instances of them and it is only minimally connected to abstract descriptions. General judgments of attachments or of types of attachments will turn out to be based on assemblages of individual judgments; therefore, by being acquainted with and judging a remarkable instance, we can change our general judgments, seeing a new way in which attachments can be valuable. To learn about attachments in general, we need to learn about particular attachments, and literature is in an ideal position to provide us with this knowledge.

The plan is as follows. In the second section, I will set up the discussion. In the third section, I will present and argue against what I call the 'classic view' of how evaluative properties of attachments relate to non-evaluative ones, while in the fourth section I will defend my view, which I will label the 'individualistic view'. In the fifth section, I will show how the view I defend renders what we learn from literature important. Finally, in the sixth section, I will discuss further implications.

2. THE SET-UP

Let's start with the sceptical view regarding the epistemic importance of the acquaintance I described in the previous chapter. A sceptic might say that what we learn from a work is *about that particular work* and about what is described in it, not about ethics in general. Regarding our analysis of *The Age of Innocence* from the fifth chapter, they can tell us: 'Yes, you have understood the novel well, with all its nuances, and realised that Ellen and Newland had such-and-such attachments, and you have evaluated them well. It was not an easy one, and not many people would have done well – congratulations! But this is just an instance of a correct application of concepts and of correct evaluations of instances and is not in any way relevant for general ethical inquiries. You have only learned something about those individual attachments, not about attachments in general.' The value of attachments, the sceptic might argue, is understood by reflecting on possible characteristics of attachments and their role in our lives, a reflection that should be done in abstraction from our engagement with novels. Even more, how could acquaintance with one particular attachment give any general knowledge? (cf. Lamarque and Olsen 1997, pp. 394-7)

Yet the intuition of many a reader is that, far from being only about that work, the understanding and judgment prompted by a work of literature reach out into the world and give us some more general ethical understanding. After reading a work by Austen, one tends to say that one has learnt something about love in general, about the value of love, not just about the love that the heroine had for her beloved. And a work by Dostoevsky might give one an understanding of faith and nihilism in general, not just of the faith and nihilism of the characters. Needless to say though, when pressing people further to say exactly what they have learnt, many of them struggle to put things into words.

The assumption behind the sceptical worry above seems to be that we *can* understand the value of attachments by reflecting, in abstraction, on what *features* might make them valuable, and, in particular, that there *are* features of attachments that are *easily connected* with value. To see why this assumption might not be as obvious as it sounds, we might think of our practices of attributing aesthetic properties to works of art. To understand what makes works of art graceful, vulgar or profound, we don't think in abstraction which non-evaluative features of artworks contribute to their being such. Indeed, as Sibley (2001) argued, there doesn't even seem to be a direct relation between descriptive properties of artworks and aesthetic properties. To understand art and why it is valuable, we need to engage with works of art, we need to acquaint ourselves with works of art and try to understand them. Basically, I want to argue that something similar happens in the case of understanding the value that attachments might have.

So, what we have to examine in this chapter is the relation between the non-evaluative features of attachments and our evaluation of them. By our 'evaluation of them', I mean our ascribing intrinsic value to them. Of course, attachments can contribute to our health or have other instrumental value, but this is not what we am interested in.

The view that would support the sceptical worry against my proposal of literary cognitivism, is what I will call the 'classic view', namely that there are a number of quantifiable features of attachments that account for their value[51]. Such a view would yield what I will call a 'smooth' transition from the non-evaluative to the evaluative: according to it, we can know which features make attachments good, and if we know how all these quantifiable non-evaluative features contribute to the value of an attachment, we are in a position to evaluate it.

Against this, I will put forward an alternative view, which proposes a non-smooth transition from the non-evaluative to the evaluative – à la Sibley (2001) – a view which owes a lot to Michael Tanner (2003). It will take the form of an 'acquaintance principle': the value of a particular

[51] I take it that a version of this view seems implicit in most writings on the topic, from Aristotle (2014) to contemporary philosophers such as Telfer (1971), Annis (1987), Thomas (1987), Cocking and Kennett (1998), Blum (2003) and Nehamas (2016). It is very rarely stated as such though, so this is a contentious matter.

attachment is revealed in that particular instance and is only minimally connected to a type of attachments or to a general non-evaluative description of attachments that applies to that attachment. From this, it will follow that tentative judgments of the value of attachments in general, or of types of attachments, such as those involved in friendship or romantic love, will be based on assemblages of individual judgments that reveal the potential of attachments in general or of a certain type of attachments. This will defuse the sceptical worry about the transition from the particular to the general, showing that we reach general judgments by judging particular instances.

3. THE CLASSIC VIEW

There are many authors who make claims about what makes love or friendship valuable, but I haven't found many detailed discussions that go into the nitty-gritty of what would make one instance valuable. Some philosophers give accounts of what might make friendship in general valuable, without going too much into details of what differentiates different instances of friendship. Others seem to tacitly assume that what is constitutive of love or friendship is also what makes them valuable. In any case, let's enumerate some frequently-encountered ideas. I will rephrase all in terms of attachments in order to make matters easier.

The Aristotelian-minded might see the value of an attachment in a certain delight in the other, especially if the other is virtuous. Note that this claim need not be interpreted as the attachment being a means to delight, which would be what is actually valuable. Rather, the delight can be a manifestation of the attachment, so it could then be interpreted as a feature of the attachment, a feature that makes it valuable. Aristotle himself talks about the goodness of perceiving and delighting in the virtue of one's friends and the goodness of benefiting them (2014, pp. 174-7). More recently, Elisabeth Telfer values the 'pleasure of the friends' company and of shared activity with someone of kindred outlook', and also claims that ‘[f]riendship makes us ‘more alive’ because it makes us *feel* more’ (1971, pp. 239-40).

Someone of a more Kantian outlook can claim that in friendship, there is a special recognition of the objective importance of the other. David Annis claims that ‘friendship is in part to be valued because it involves recognising the deep value of the person’ (1987, p. 351). Again, we can see this recognition as a manifestation of the attachment and hence as an aspect of the attachment that makes it valuable, not merely as something that the attachment is a means to.

More down-to-earth, Laurence Thomas talks of the importance of the ‘enormous bond of mutual trust between … friends’, a bond that ‘is cemented by equal self-disclosure and, for that very reason, is a sign of the very special regard which each has for one another’ (1987, p. 217).

Dean Cocking and Jeanette Kennett talk about the importance of reacting to the particular traits of the other, to their view of us, and of our changing in light of this interaction, essentially of being ‘receptive to being directed and interpreted and so in these ways drawn by the other’ (1998, p. 503).

Lastly, there are some writers on the topic who are very liberal and inclusive in the good-making features they ascribe to friendship. Lawrence Blum lists:

> deep concern, involvement, commitment, care, loyalty, intimacy, and other virtues, sentiments, and qualities taken to characterize worthy instances of personal relations (Blum 2003, p. 512),

while Alexander Nehamas claims that:

> The benefits of friendship are many. The love friendship provokes gives depth and color to life; the loyalty it inspires erodes the barriers of selfishness. It provides companionship and a safety net when we are in various kinds of trouble; it offers sympathy for our misfortunes, discretion for our secrets, encouragement for our efforts. (Nehamas 2006, p. 187)

Now, most of these writers do not expand too much on how exactly to cash out these good-making features of attachments (or of specific kinds of attachments or attachment-based interactions). Some of these features might seem just general goods of life that attachments might be instrumental to. But I take it that at least some of these writers think the attachments themselves are good in virtue of some of these characteristics. Now, if we construe these features as themselves evaluative ('loyalty' might be an example), we do not really have so much of an explanation of what makes attachments valuable, for we should now ask how these evaluative features themselves relate to the non-evaluative ones. So, we should try to construe them as non-evaluative features of the attachments, that account for their value. Also, it is natural to construe them in some way as quantifiable. For instance, there might be more or less trust in a friendship, or one can delight more or less in the virtue of the person one is attached to, or one can recognise more or less the objective value of the other.

What I will call the 'classic view' is the theory that the value of attachments is dependent in a simple way on some of these quantifiable properties. I start with a simple version of it, not so much because I think people actually hold it, but because it serves as an introduction to the holistic version that will be an improvement on it.

3.1. THE SIMPLE VERSION

According to this simple version of the classic view, there are quantifiable non-evaluative properties $p_1,p_2,\ldots,p_n$ and $q_1,q_2,\ldots,q_m$ such that an attachment is valuable insofar as it has $p_1,p_2,\ldots,p_n$ and does not have $q_1,q_2,\ldots,q_m$. As discussed above, possible candidates for the p_is are trust, delight in the goodness of the other, being receptive to their opinions, delighting in their happiness and willing to contribute to it and so on. Of course, one might claim that there are only p_is and no q_is, but if we are to think of candidates for the q_is, one might come up with the other person being immoral[52].

[52] I realise that some of the properties that I have listed might be taken to be evaluative. Some terms, such as 'delight', are ambiguous, as they are sometimes used descriptively and sometimes with an evaluative tint. I want to use the purely descriptive version of them. Also, some other of these properties contain evaluative terms: for instance, the

Anyway, the crude idea is that for each of the properties, an attachment A gets a score – call it $p_i(A)$ – which represents how much it instantiates the property. Then, we can compute the value of A by adding all the $p_i(A)$s and subtracting all the $q_i(A)$s.

This simple version of the classic view thus posits a very smooth transition from the non-evaluative to the evaluative: knowing a handful of non-evaluative features of the attachment (the scores it has on the properties, that is, all the $p_i(A)$s and $q_i(A)$s) puts one in a position to judge the value of that attachment without knowing anything substantial about further details of the attachment.

According to the simple version, is there any role for acquaintance with particular instances of attachments? Well, not really. Of course, in order to understand how attachments can be valuable, one needs to understand the properties p_is and q_is and why they are good-making features. These might actually be properties that appear in other areas of life, so need not necessarily be grasped from attachments. If they do not appear elsewhere and are only features of attachments, then some acquaintance with attachments might be necessary in order to understand what these properties are and why they are good/bad-making features, but this is a minimal form of acquaintance. Moreover, the examples that would serve to grasp these properties and why they are good-making ones would be very bare examples, stripped down of all details that are not relevant, in order for one to easily discern the relevant property. They would be like the typical examples used in thought experiments in analytic philosophy rather than like the intricate examples depicted in novels.

Once one has understood what the p_is and q_is are and why attachments are (dis)valuable in virtue of them, it seems that there is nothing general left to be understood about the value of attachments. Indeed, when being acquainted with a new instance, one just goes through the motions of calculating the scores and that would be it.

property of an attachment that it is 'an attachment to an immoral person'. However, note that as a description of an attachment, this is not evaluative, just as 'talking to a moral person' is a non-evaluative description of an action – such an action might be good or bad.

Of course, it might sometimes be difficult to see how much one of the properties is instantiated in a certain attachment. But this difficulty relates to the non-evaluative description of it, and given that these properties are quantifiable, finding out how much the property is instantiated does not amount to a new bit of understanding of the value of the attachment. (Compare with the following: to calculate the density of a country, one needs to calculate the surface, and this might be difficult if the country has a peculiar shape. But doing this does not provide one with any novel understanding of density.)

Now, someone might protest that I should modify this version of the classic view by allowing the p_is and q_is to be thick evaluative properties rather than non-evaluative ones – perhaps one might think of properties like 'delicate', 'vulgar' etc. As suggested above, I don't object to such a version, but it is not what we are looking for. The question that we are asking is how the evaluative arises from the non-evaluative, so if the p_is and q_is were evaluative, then the question would become, how do we judge the extent to which an attachment is p_1 (e.g. 'delicate')? And the same discussion would ensue at that level.

People might consider this simple version of the classic view naïve or simplistic, and even if they are attracted to the idea that there are various non-evaluative properties that make attachments more valuable and some that make them less valuable, they might suggest that one good-making feature need not always count towards the goodness of an attachment. For instance, someone might claim that trust is good only if the other person is moral or that a desire to contribute to their happiness is good only if it is coupled with a desire to understand them. This points to the more sophisticated, holistic version of the classic view. I won't discuss other reasons why I reject the simple version, as all the arguments against the holistic version will apply to this one as well.

3.2. THE HOLISTIC VERSION

We can thus put forward a more sophisticated version of the classic view, similar to Jonathan Dancy's holism about reasons (2004, esp. pp. 3-12, pp.

73-78). The idea is that we still have the list of properties $p_1,\ldots,p_n$ and q_1, $\ldots,q_m$ such that the p_is add to the value of an attachment and the q_is subtract from it. And the total value is still the sum of contributions of the p_is from which we subtract the contributions of the q_is.

However, here is the key difference: the contribution of each p_i to the value of the attachment A – call it $F_i(A)$ – does not depend solely on how much p_i is instantiated in the attachment, but on other factors as well. In other words, $F_i(A)$ is not always equal to $p_i(A)$ and might even be negative. And similarly the contribution of q_i – call it $G_i(A)$ – is not always equal to $q_i(A)$. To use the example from before, perhaps trust is good if the other person is moral and bad if not. And one can go even further and say that if the other person is immoral, yet the attachment has a kind of sacrificial nature (as was Sonya's attachment to Raskolnikov in Dostoevsky's *Crime and punishment*), then it is again good ('you have to believe in him to save him from himself'). Adding more factors in, we can argue that if the other person is immoral, the attachment has a kind of sacrificial nature, but this sacrificial nature is too high-minded, principled and perhaps self-satisfied (as again Dostoevsky makes us think is Katerina Ivanovna's attachment to Dmitry Karamazov in *Brothers Karamazov*), then it is bad. And we can go on like this. Importantly though, in all these case, it is still the *trust* that contributes to the goodness, or lack thereof, of the attachment, it is still trust that makes the attachment good or bad. Even though there are other factors involved, in the examples described the good-making property is trust and the remaining factors are just enablers and intensifiers, that allow (or not) trust to contribute, more or less, to the goodness of the attachment[53].

The upshot is that even if we know all properties which *might* contribute to the value of an attachment, *if* and *how much* they contribute to an actual instance can be very context dependent, and there might be no general principles that codify this. Any attempt at systematic codifying could have exceptions. This means that the holistic version posits a less

[53] See Dancy (2004, pp. 38-52) for a detailed view, similar to this one but in the realm of moral reasoning, of the difference between a reason, an enabler and an intensifier. Väyrynen (2006) talks about 'hedged' moral principles that state what features are moral reasons, principles that can have exceptions, yet which form the basis of moral reasoning. The view regarding the value of attachments that I am presenting here is an adaptation of some of these ideas.

smooth transition from the non-evaluative to the evaluative than the simple version: no general description of an attachment can license someone to judge that attachment in more than probabilistic terms. Knowing how much an attachment A instantiates p_i and q_i, that is, $p_i(A)$ and $q_i(A)$, does not immediately put one in a position to judge the attachment. However, knowing how much all relevant properties are instantiated in the attachment, with an 'and there is no other relevant factor' clause added, would put one in a position to judge the attachment. In this sense, the transition from the non-evaluative to the evaluative is still somewhat smooth.

Does acquaintance play any role according to the holistic version? It does play a role in judging individual instances. Indeed, the natural way to judge an instance of an attachment is to be acquainted with it, to see how much the good-making properties are instantiated, to observe other relevant factors which might change the contribution of the good-making properties and, *importantly*, to check that there are no further relevant factors.

However, although acquaintance is needed to judge instances, the general understanding of the value of attachments is not gained through those instances. Indeed, even though, unlike in the simple version, in the holistic version the relevant properties are wayward in their contribution to the value of the attachment, it is still them that *make* the attachment valuable or not. Hence, to understand why attachments in general are valuable, one needs to understand the p_is and the q_is and why they are good/bad-making features of attachments.

A proponent of the holistic view can reply that there are two ways in which acquaintance with various instances can lead to some general form of understanding.

First, one can in this way formulate more and more precise principles that, even if they might have exceptions, apply with an 'if there is no other relevant factor' clause. Indeed, given the way we have built the theory, the basic principles would be of the form: 'An attachment is better to the extent that it has property p_i, unless there is some other relevant factor which invalidates or changes the contribution of p_i.' We can for instance think of p_i as being 'trust'. However, as discussed before, there might also be a principle of the form: 'If the person one is attached to is immoral, the attachment is worse to the extent that it involves trust, if there is no other relevant factor.' And we can devise more and more complicated principles in this way. These would perhaps constitute a form of understanding that is derived from acquaintance with more and more complicated instances.

One thing to say about this move is that the kind of acquaintance that it involves is not the kind of acquaintance that we get in literature. Indeed, formulating a new principle would essentially involve finding a type of exception to a previous principle, and for this exception to be clear, it should instantiate little beyond what makes it an exception, as any other aspect might detract us from what makes it an exception. To come up with the principle that trust is bad in an attachment to an immoral person, the example from which to derive this principle would involve just the trust and the immorality of the other. It would again be the kind of example that is used in many thought experiments in analytic philosophy. And indeed, it would hardly be the type of acquaintance presented in the previous chapter.

Another thing to say is that it is not clear to me how deep the understanding involved in formulating finer and finer principles can be. After all, what we are doing is finding exceptions, exceptions to the exceptions and so on. It's true that we get a better picture of how these factors combine in making attachments good or bad, but we should remember that what *makes* them good or bad are the properties p_i and q_i. The fine-grained principles are just applications of this general fact.

We can now move on to the second kind of understanding gained from acquaintance with instances that the proponent of the holistic version might appeal to. Such a proponent would presumably contend that it is difficult for someone to judge an individual instance of an attachment, to see how much the properties p_i and q_i are instantiated and which other factors influence their contributions to the value of the attachment. There might be a lot of contextual information and one might not know which parts of this information they should take into consideration. In virtue of this, we might hope that there is a way to train oneself to navigate the contextual information and make the right judgment. And one possible way to learn this would be, adapting the ideas from Martha Nussbaum (1990), just to be acquainted and taught to evaluate instructive instances. One just trains one's sensibility and intuition such that one becomes a better judge. It would thus follow that acquaintance with rich instances might be helpful in training oneself to be a better judge.

It would be a matter of dispute whether this is the best way to train oneself. Someone might claim that the best way to go is to formulate more and more fine-grained principles. I take it that this would be to some extent an empirical matter that would also depend on how one fills in the details of how the properties and exceptions actually tend to look.

But, even if we accept that this kind of acquaintance is needed to train oneself, it still seems a relatively minor role that acquaintance plays. First of all, it is a contingent matter that one needs training – after all, perhaps people could be born with the relevant capacity to judge. Second, and more plausibly, once one has trained oneself, one can forget all the instances that helped one train oneself, just as after learning to ride a bicycle, one can forget all the exercises one has done in order to learn. It's just the know-how that matters.

Having explored the implications of the holistic version, I will now move on to explain why I think it is not right. None of the arguments against it will be decisive in itself, but I hope that together they will make a persuasive case against it. Moreover, after presenting my own view, I hope it comes across as a better candidate.

The first worry would be that it seems possible that a small difference in the non-evaluative description of the manifestations of an attachment can make a significant difference in the value of the attachment. Indeed, we can return to the example of the two mothers from the first chapter:

> *Mother A*: Mother A is very involved in her daughter's ballet lessons. They both hope that she becomes a good ballerina and devote a lot of time to this. The mother accompanies the daughter to all the shows, takes her to the doctor to preempt any possible medical problems and feels joy when the daughter progresses. Sadly, when the daughter turns 16, her body develops such that it is impossible for her to become a world-

> class ballerina. The mother suffers with the daughter, yet encourages her: 'Now you can pursue your interest in philosophy, which is decent as an alternative.'
>
> *Mother B*: Mother B, exactly as mother A, is very involved in her daughter's ballet lessons. They both hope that she becomes a good ballerina and devote a lot of time to this. The mother accompanies the daughter to all the shows, takes her to the doctor to preempt any possible problems and feels joy when the daughter progresses. Sadly, when the daughter turns 16, her body develops such that it is impossible for her to become a world-class ballerina. The mother gets very annoyed and says to her daughter: 'I've wasted my time with you' (and really believes that, not only in that moment).

The manifestations of the attachments of the two mothers are very similar, in that the sets of emotions that they feel towards the progression of the daughters are very similar. Mother B's reaction at the end might come across as shocking, but we should keep in mind that from a strictly quantitative point of view, it is just one reaction amongst many. Yet this one reaction makes us judge the two attachments differently. I take it that while mother A's attachment is a very valuable one, mother B's is not. Can the proponent of the classic view account for this change in value? If the properties p_i and q_i are descriptive, quantitative properties, then it seems that how much they are instantiated in the two attachments should be very similar, given that the manifestations of attachments are very similar. It follows that, to account for the change of value within the framework of the classic view, one needs to postulate that there is a change in the factors which modify the contribution of the p_is and the q_is to the value of the attachment. The proponent of the classic view would have to argue, for example, that the contribution to the value of the attachment of some of the p_is is invalidated by other (non-evaluative) features of the attachment. For instance, the care for her daughter's projects has value for mother A's attachment, but, in the case of mother B, her final reaction prevents this care to contribute to the value of the attachment (or even makes it contribute negatively).

The problem with this strategy is that it starts to look a bit like a subterfuge to accommodate any possible counter-example. I take it that the idea behind the classic view is that we should be able to make sense of and understand the transition between the non-evaluative and evaluative. But the strategy just described above implies that while the non-evaluative good/bad-making properties of attachments (the p_is and the q_is) are easy to understand, grasp and control, the other features, that influence the contribution of the good/bad-making properties, are not really easy to control. This is because the strategy implies that a very small change in the manifestations of an attachment can lead to a significant change in these features. If the proponent of the classic view wanted to make that view more solid, they could claim that not only the properties p_i and q_i that make attachments (dis-)valuable are quantifiable descriptive properties, but also the properties that invalidate or intensify the contribution of the p_is and q_is. But if they make that move, my counter-example goes through. This is because all quantifiable descriptive properties of the attachments of mothers A and B should be similar, including those that invalidate or intensify the contribution of the p_is and q_is. This means that the contributions of the p_is and q_is, that is, $F_i(A)$ and $G_i(A)$, respectively, would be similar, meaning that the values of the two attachments are similar.

The second objection that I have to the classic view is more powerful. Essentially, the worry is that the view makes judgments of attachments too similar to practical judgments of what to do, and my contention is that the two types of judgments are different. The various non-evaluative features of attachments are seen as similar to reasons in practical deliberation, which are weighed together and against each other. The classic case is that of breaking a promise in order to help someone in need: the fact that the action constitutes the breaking of a promise counts against it, while its being a case of helping counts for it – we just need to see which reason is more powerful. But in the case of various non-evaluative features of an attachment that are supposed to guide our judgment, they do not merely compete against each other, but rather *illuminate* each other. We should keep in mind that, as I have argued in chapter 1, all the manifestations of an attachment that we put together in order to judge that attachment are manifestations of the same mental state, the attachment itself, and should be regarded as such. For instance, the last remark of mother B helps us see all the manifestations of the attachment in a new light. We start seeing her previous actions not as devoted, as we might have been tempted to think, but controlling, as revealed by the last reaction. In a sense, we judge the attachment as a whole, we don't judge individual parts and then add them up. This is unlike the case of practical reason, in which we *do* often judge individual aspects and then add them up – the fact that an action constitutes a breaking of a promise and the fact that it is an act of helping do not illuminate each other and can be judged separately without any problem. Anticipating what I am going to argue in the next sections, we can see that judgments of attachments in light of various features of them are more like judgments of artworks in light of various features of them than like judgments of actions in light of various reasons for or against them.

The last problem is that there are many idiosyncrasies in attachments, idiosyncrasies that plausibly contribute to the value of those attachments. The common history with the person one is attached to might have created a special significance attributed to activities done together and the peculiarities of the other. For instance, in *The age of innocence*, we might think of how the books and flowers that usually decorate Ellen's abode, in particular azaleas, come to be associated by Newland with her individuality in the world of New York. When observing how the flowers are arranged in her drawing room, he gets the feeling that the place is special; and, when he finally imagines her in her Parisian flat, he can't help but imagine the room dominated by flowers. This is what we can call an idiosyncrasy of the attachment. These idiosyncrasies seem to contribute to its value, and it is not clear how the classic view can account for this.

In reply to this, the proponent of the classic view could try to group possible idiosyncrasies as a broader type of good-making features. Perhaps they might claim that a good-making feature might consist in associating a special meaning to the other's eccentricities. But it is strange to say that any meaning associated to any eccentricity is good-making. Moreover, there seems to be something hubristic in thinking that one might give a short description of all ways in which idiosyncrasies in attachments *could* be good, in general.

To conclude, I have put pressure on the classic view, even in its holistic version. Given that what I called the 'classic view' is more a type of theory than a theory, with many details left to be filled in – for instance, what the good-making properties are and what other properties influence their contribution – its plausibility depends on whether its proponents manage to fill in the details in such a way as to make it work. It's hard to settle the matter at this level of abstraction, but I hope the theory that I will put forward will make better sense of the phenomena we are observing and trying to accommodate.

4. THE INDIVIDUALISTIC VIEW

Following some of the worries that I have raised about the previous view, I can now propose my own view, which I will call 'the individualistic view'. In its shortest formulation, it could be expressed in the following 'acquaintance principle'[54]:

> ACQUAINTANCE PRINCIPLE: The value of an instance of an attachment is revealed in that instance. It is only minimally connected to a concept (e.g. friendship) or to a general abstract description that applies to that instance.

So, in order to judge an instance of an attachment, we shouldn't ask which features it has that make attachments in general valuable, but rather be

[54] Note that I am using the term 'acquaintance principle in an unusual way. In aesthetics, the term 'acquaintance principle' is applied to several loosely related ideas, including the following: that a judgment of a work of art should be based on first-hand perceptual acquaintance with the work, that one should not base one's judgment on the testimony of others (either because there is something inherently wrong with that, or because it is highly unreliable), that an aesthetic judgment should involve a kind of personal involvement in it, etc. (Budd 2003, Meskin 2004).

acquainted with it (in the way described in chapter 5) and judge that instance[55].

Yet we do ask questions like ‘why is love valuable?’, questions which seem to expect a general response, one that applies to all instances of love. According to my view, this kind of question is slightly misleading. We can indeed give an answer to it, but that answer will be derived from the individual judgments that we have made. These individual judgments put together offer a glimpse of ways in which love can be valuable. We thus have the next principle, as a consequence of the first:

> GENERAL JUDGMENTS FROM SETS OF INDIVIDUAL JUDGMENTS: When judging a type of attachment (romantic love, loyalty, family love, etc.), we actually derive this general judgment from a set of individual judgments that we have made. The individual judgments point to the potential of that type of attachment.

This might all sound rather peculiar and contrary to the established way of doing ethics, so I think that a useful way to make this more palatable is to highlight that this is similar to how judgments of art work. In effect, I am arguing that judgments of attachments work somewhat like judgments of art.

The acquaintance principle, formulated above, exhibits an analogous structure to Frank Sibley’s characterisation of aesthetic judgments (2001, esp. pp. 3-13). According to Sibley, no non-aesthetic description of an object (e.g. ‘the vase is curved, so-and-so tall, sky blue etc.’) can license the application of an aesthetic predicate (e.g. graceful, delicate, intoxicating, vulgar etc.). The application of such a predicate is typically the result of acquaintance with the property in the object.

[55] A hint towards this view is found in Michael Tanner: ‘[I]t is characteristic of certain moral qualities that they too can only be ascribed to someone on the basis of first-hand experience of him and his behaviour. This fact is concealed because the predicates designating the qualities often are used, and correctly, in a way that *can* be inferred from a description; I am claiming only that they are not always used in this sense, and that is especially characteristic of their use in what might be called ‘morally creative’ contexts that they are not.’ (2003, pp. 29-30)

Also, as explained in chapter 5, the notion of acquaintance we are working with is somewhat different from the one involved in art. When seeing a painting, there is a sense in which we see it all (even if we might not observe it in the right way, with the right expectation, knowledge etc.) In this narrower sense, we can say that one can be fully acquainted with a painting. In the case of an attachment, acquaintance cannot be complete, since one cannot imagine all real and possible emotional manifestations of that attachment. However, there is a point when knowing enough manifestations of the attachment gives one a good enough understanding about that attachment to be able to pass a judgment.

There is, of course, a possibility that one is misled in making a judgment of value of an attachment. For one, one can be misled about the attachment one judges by not being well enough acquainted with it. In such a case, getting to know a new manifestation – a manifestation that one hasn't been aware of – might change the judgment. Yet one might also be well acquainted with an attachment, that is, as acquainted as needed in order to judge it well, but pass a wrong judgment. My claim is only that in some cases partial acquaintance is enough for passing a correct judgment and a that correct judgment is passed in some of these cases.

I should say something about the question of how we recognise the value in an attachment, that is, what is supposed to happen in the mind of someone who recognises the value in one instance. According to some theories, one recognises value by admiring the person that instantiates the value, that is, by having an emotion (Deonna and Teroni 2012, Zagzebski 2017). According to others, one is first struck by value, and the emotion is a reaction to that (Müller 2017). I do not want to commit myself to an answer to this question, but, especially given some of the things argued in the second chapter, it is more natural (though not obligatory) to go towards a version of the second view. Under such a view, an emotion of admiration might signal not so much that one has grasped the value, but that the value that one has grasped matters in some way to one. One can also grasp the value without really caring about it.

Regarding the principle I proposed for general judgments of attachments, or judgments of types of attachments, this is again similar to what happens in art. Indeed, according to my view, asking ‘why and when is friendship valuable?’ and expecting a short answer is like asking ‘why and when is a painting valuable?’ and expecting a short answer[56]. To find out what makes paintings valuable, one needs to engage with paintings, in particular with good ones, to see them, think about them, judge them, compare them, etc. Once one has done this, one might indulge in some general reflections on painting, but these have to be based on the individual judgments that one has made. These individual judgments have shown ways in which paintings *can* be good. Someone who has never seen a painting, or who has only seen a relative’s watercolours would presumably have little idea of the ways in which paintings can be good and how good they can be.

Having explained the view, I will now proceed to arguing that it is better than the classic view.

The first point, perhaps a moot one, is that we shouldn’t start from the classic view as the default position in this debate; in fact, there is no special burden of proof on someone proposing a view like mine. It is not clear that our evaluative practices fit the classic view better than the individualistic one. Indeed, when we spontaneously admire attachments in daily life or in novels, it is not obvious that we admire them because we find some quantifiable features that appear in high degree in those attachments.

We might give some kind of motivation for our evaluation – for instance, we might say ‘Just pay attention to how she listens to her child!’ But this does not imply that someone offering this kind of motivation thinks that the more carefully one listens to one’s child, the better the attachment. Indeed, we might think of a parallel with art: when pointing to the triangular composition when motivating our judgment of a painting by Rafael, we do not thereby claim or imply that a triangular composition is something that adds to the value of paintings in general. Instead, we might be just suggesting to our interlocutor to see the painting in a new light, by focusing on this composition, with the hope that they will experience it in a more satisfying fashion and hence regard it as more valuable than they originally did. Similarly, when pointing to a feature of an attachment as an indication

[56] Of course, one might claim that paintings are good because they give happiness or have some other kind of instrumental value. But I am only referring to intrinsic value.

that the attachment is worthy, what we might aim for is just that the interlocutor focuses on some aspects and, therefore, sees the attachment in a new light. As mentioned before, given that manifestations of an attachment stem from a common source, namely, the attachment itself, we should see them together, since one of them may illuminate the others. If we observe that Jane is very attentive to what her young child says and follows his train of thought very carefully, we might come to reinterpret her involvement in helping him learn. Indeed, that involvement would come across not only as a form of caring, but also as a form of living his childhood with him, without any further goal of helping him develop. All this serves to see the value of that attachment.

The second reason for accepting the individualistic view is that it can easily account for the phenomenon, mentioned above, that a very small difference in the manifestations of an attachment can lead to a significant difference in value, as in the case of mothers A and B. This is again very similar to the case of art, in which a small difference – for instance a stroke of paint in a painting, or a word in a poem – can lead to a significant difference in value. If I am right in claiming that the value of an attachment is grasped in the instance and that the transition between the descriptive and the evaluative is not – as the classic view proposes – smooth, then there is no reason why a small difference in the manifestations of an attachment should not lead to a large difference in value.

The third reason is that the view can easily account for the value of idiosyncrasies. If the value of attachments is grasped in each instance, this means that we might be surprised by finding value connected to some idiosyncrasies of attachments, idiosyncrasies that we might not have even thought could be of relevance. In *The Age of Innocence*, we might think of how the sacrifice Ellen and Newland made is embodied in their attachments, as revealed at the end of the book, and also of how Ellen's abode, with her flowers and books, is given a special significance in Newland's memory.

This connects nicely with a fourth reason. As argued in the first chapter, the beginning of an attachment is a kind of creation, in the sense that a new drive appears, a drive that, by definition, is not a manifestation of another mental state. This means that the attachment is not a manifestation of a previously held aim. In particular, it follows that it is not a manifestation of an aim to have the characteristics that the classic view proposes make attachments good. Similarly, the development of an attachment is also not guided by such an aim. Instead, it is a response, motivated by the attachment itself, to events lived together with the person one is attached to and to new discoveries about them. The other person features essentially in the attachment, and their traits and common history gain some special significance in the attachment. Both the creation and the development of the attachment are thus focused on the person one is attached to and are not motivated by abstract goals. It would thus be more natural for evaluations of attachments to concentrate on these particular responses to this particular person. Judging them by abstract standards that would apply to all attachments would not seem consistent to the way the attachment is formed and develops. The view that I am proposing does more justice to the formation of the attachment: the focus of the person judging would be as particularised as the focus of the person having the attachment.

I hope to have shown that my view fits better than the classic view with many aspects of attachments: how they are created, their unity, our practices of appreciating them and the fact that subtle differences in manifestation can signify significant differences in value.

5. REMARKABLE ATTACHMENTS AND THE ROLE OF LITERATURE

We are now in a good position to see the role of being acquainted with remarkable attachments, in particular with those that are depicted in great works of literature.

Given that general judgments of attachments, or of kinds of attachments, are based on assemblages of individual judgments, any new individual judgment changes those general judgments. A new individual judgment might change the general ones only a bit, if it fits in very well with the previously made ones. But if one is acquainted, as we often are in very good works of literature, with a remarkable attachment that does not fit too neatly with the ones previously made, the general judgments might change significantly. Let's develop this point.

The first thing that remarkable attachments can do is to show something about the potential of attachments. To take the simplest of examples, suppose that the attachments of people around us are not in any way remarkable or profound. We then read a novel (e.g. *Middlemarch*) and are acquainted, for the first time, with a profound attachment. Once we are acquainted with this attachment, we realise, silly as this might sound, that attachments *can* be so remarkable, or that this particular type of attachment can be so remarkable. Before that, we just didn't know, and there was no way to just deduce that they can be so.

Also, a remarkable attachment, such as those depicted in *The Age of Innocence*, can show us new ways in which attachments can be profound or valuable. Indeed, let's remember our discussion from the previous chapter that, at the end of the novel, the attachment that Newland had to Ellen was such that it prevented him from pursuing a pleasant, yet ultimately superficial encounter. I take it that by being acquainted with it, we realise that an attachment can be valuable not so much in spite of this kind of refusal, but with this refusal as an integral part of what makes it valuable.

A remarkable attachment can also help us see the value of lesser attachments. It is a fairly natural assumption that attachments that do not impel one to spend some time with the person one is attached to cannot be too powerful, or indeed valuable. Reading a novel like *The Age of Innocence* can not only change our mind about this, but it can lead us to observe the value in similar attachments of people around us. Before, we might have thought very little of friends that do not bother to see each other often, assuming that their attachments cannot be particularly valuable. Now, we might gain a better understanding of them, see them in a new light and grasp some value that has escaped us so far. It's not so much that we imagine some emotions that are manifestations of these attachments that we wouldn't have imagined before. Rather, we see all the emotions that we imagine and all the behaviour that we observe in a new light, for instance as expressing an avoidance of superficial encounters.

Lastly, even though this goes well beyond the purpose of this chapter, remarkable attachments can influence the formation of new attachments, just like ground-breaking works of art influence many other artists. Each of Picasso's *Les Demoiselles d'Avignon*, Wagner's *Ring* tetralogy and Truffaut's *Les 400 coups* have deeply influenced their respective art form (and not only). This sort of influence can happen in attachments, even if the influence might be smaller and the hierarchies less clear-cut. But clearly, this 'influence' is anything but imitation, and to elucidate it is well beyond my purposes here; indeed, it is an immensely difficult task.

6. THE OPEN-ENDED NATURE OF JUDGMENTS OF ATTACHMENTS

I want to end by discussing some implications of the view that I have proposed, implications that seem to me right, giving further credence to the view, but which others might look at differently. Anyway, it's good to have the cards on the table.

First, as mentioned before, general judgments of attachments are by nature tentative, open to further revisions. This is not so much because of some limitations of human judgment, whatever such limitations might be, but because general judgments are based on assemblages of individual judgments, therefore constricted by what the judger has been acquainted with. Any new acquaintance with an attachment might change the overall judgment. To take the example I've used so far, one might believe that attachments that do not impel the agent to spend time with the person they are attached to cannot be particularly profound. This judgment would be based on the attachments one is acquainted with, perhaps those of people around one. If one then reads *The Age of Innocence*, one can see that the attachment of Newland to Ellen is profound, yet does not impel him to go up to her; accordingly, one might change one's general judgment of attachments on this aspect.

Second, following on from the first point, the theory I have proposed implies that there is, or at least can be, a form of progress in history. Indeed, one’s general judgments are limited by the attachments one is acquainted with in real life and in literature. As time passes, more examples appear, either in real life or in literature, and, therefore, we might be better placed to pass a general judgment. It follows that our judgments might be better than those of our stone age cousins or of, say, the ancient Greeks that we otherwise admire. This would be not in virtue of some superior sensitivity, but simply in virtue of there being, as it were, more material to judge. It’s true that some attachments or types of attachments might be lost in history and that we might be unable to see the value of some attachments of people long ago, but I think we can still hope that we are in a better position than our predecessors. I realise that I start to sound rather Whiggish, which is far from my intention, but I believe there is a presumption that a well-functioning culture, that keeps alive some remarkable instances of attachments and creates new ones, is going towards better general judgments.

Third, as an ontogenetic version of the previous observation, there is a sense in which one’s general judgments of attachments might get better with age. Again, this is not because one might get wiser and one’s sensibility more refined with time, but because one has been acquainted with a larger number of relevant examples.

Lastly, I think the theory that I have defended allows for objective judgments while also allowing for some kind of diversity. In common speech, when someone is deemed ethically good, it is sometimes assumed that the judgment involves some kind of normativity to the effect that completely different people from the praised one are not quite good. The view I proposed does not have that implication. Even if we judge an attachment to be very valuable, this does not in any way imply that value cannot be realised in very different attachments. Yet neither does it imply that there cannot be any judgments of comparative value, that anything goes.

7. CONCLUSION

I have argued that the value of attachments is revealed in the instances and that our general judgments as to the value of attachments are based on assemblages of individual judgments. Works of literature can give us knowledge of the remarkable attachments that are depicted and, as a consequence, improve our general judgments of attachments. This is a significant cognitive gain.

CONCLUSION

At the end, I could usefully summarise the main claims that I have argued for in this thesis:

1. An attachment to a person (or entity) is not a mental disposition, but a mental state that manifests itself in a mental disposition.
2. This mental state is a drive that encapsulates the importance the other person has for the agent. It is an intentional, standing mental state, that manifests itself in emotions and that is not a manifestation of any other mental state.
3. Emotions are bodily attitudes directed at a content taken from another mental state (via Deonna and Teroni), that make courses of action intelligible in terms of what the agent cares about.
4. Emotions that are manifestations of attachments do not have fittingness conditions.
5. Imagining an emotion is forming a thick meta-representation of that emotion, which gives the imaginer access to its phenomenology without having the emotion.
6. Literature can provide experiential knowledge of various emotions by drawing us into imagining them, that is, into forming thick meta-representations of them. This experiential knowledge is not necessarily worse than the one gained from having emotions.
7. Experiential knowledge of an emotion E consists in being able to imagine E and recognise E when having it (via David Lewis), but not in the ability to remember it.
8. Literature, novels in particular, can acquaint us with attachments of literary characters, attachments that are in various way remarkable, and is particularly well placed to do that.
9. The value of a particular attachment is revealed in that instance. It is only minimally connected to a type of attachments to which that instance belongs or to a general description that applies to it.

10. A general judgment of a type of attachments is derived from an assemblage of individual judgments instances that reveal the potential of that type of attachment. A new individual judgment, such as that of an attachment from literature, can change significantly a general judgment.

This is a large set of claims, covering many areas of philosophy and touching upon many topics, and it is inevitable that a project like this one raises as many questions as it has attempted to answer. While I was aware of these questions and gave them some thought, I decided to stick to what was relevant for the main stream of the thesis. Nonetheless, it would be just appropriate to point to some of these questions here.

First of all, there is the age-old debate about the conflict between attachments and impartial morality, and any new view of attachments (or of impersonal morality) is bound to give a new twist to this debate. Given the theory that I've put forward, it might seem that when there is a conflict of this kind, say between acting for a friend and performing a public duty, the two actions are motivated by completely different drives, and hence it is just a matter of which of them pushes harder. Of course, this is one possible conclusion, and an eminently plausible one, but one might still wonder whether there is something more to be said.

A second natural question is to what extent my theory about what we learn from fiction regarding attachments could be extended to other kinds of drives. For instance, we might think of interests. An interest in, say, physics might also consist in a drive, and people might be interested in physics in very different ways – something which might not be obvious at an academic conference, in which the participants talk mostly about the content of their views, not about what motivates them to wake up in the morning and go to the office or to the laboratory. Or perhaps there might be a sense in which one's interest in physics is more similar to a friend's interest in stamps, than in a colleague's interest in physics. It could be that a theory of interests works similarly to my theory of attachments, with something like the principle of acquaintance that I defended in the last chapter applying to this area as well. Literature has been somewhat less concerned with interests in physics, let alone stamps, than with romantic love, but we shouldn't assume that there is no fertile ground in this direction.

Third, I have concerned myself solely with literature (and novels in particular), and not with other narrative arts, and it is legitimate to wonder why. In particular, one might wonder why I do not discuss cinema as well. Film is also a narrative form of art, and many films could be just as good candidates as novels for learning something from them. It's true that some films might fit the theory that I have presented, but I still think there is an important distinction between literature and cinema, that makes only the former better suited to the view I presented here. The difference is that literature, unlike cinema, can do better in 'getting into the heads of characters'. Despite some techniques, such as voices from the off and filming from the character's perspective, I cannot help but see cinema as inevitably *observing* a story from the outside. The fact that it can *observe* rather than *tell* a story might put it in a better position for other projects, that might be equally important as the one described in this thesis: for instance, I think it might be in a better position than literature to problematise our very capacity to understand other people, to know what motivates them, what their drives are and so on; in other words, cinema might show the limits of the optimism that has permeated this thesis. For instance, each of Chantal Ackerman's *Jeanne Dielman* (1975), or Cristi Puiu's *Aurora* (2010) follows very closely a protagonist, who ends up committing murder. In each of them, we start guessing what might happen in their mind, but it eludes us, and, at the end, we are more convinced of the difficulty of understanding people than of anything else. Even if these films are as different as possible from the novels I've discussed throughout the thesis, I think they are equally humanistic in their placing a great importance on understanding people, even if they come from a different angle and perhaps with a different answer.

I left for last the question that, after all I have argued, seems to me most pressing, yet also most difficult. This is the question of how we build our own attachments in light of the attachments that we admire. It seems clear that we do, for our attachments often bear a resemblance to those that we admire in people around us or in fiction. Even more broadly, we might say that our attachments are influenced by the culture we are part of. But what exactly is the process in which this influence takes place? One might be tempted to say that it's a form of emulation, but 'emulation' is a dangerous word. To see why this is, let's think that I admire someone's attachment and try to have a similar one. If I imitate the model's behaviour, and force myself to feel vaguely similar emotions, then I am not having an attachment, I am just, well, aping someone. As argued, forming an attachment involves a direct response to the other person and therefore cannot be a means to being like someone else. Yet in this direct response, it seems that there is the influence of all the models, positive or negative, that one has had, from life or from literature. My hunch is that forming an attachment in light of all the attachments we admire is similar to an artist creating some new work in light of all the influences she's had. But when I start making such sweeping statements, I might just as well end here.

BIBLIOGRAPHY

Algoe, Sara B. & Jonathan Haidt. 2009. Witnessing Excellence in Action: the 'Other-praising' Emotions of Elevation, Gratitude, and Admiration. *Journal of Positive Psychology* 4 (2): 105-127.

Annis, David B. 1987. The Meaning, Value, and Duties of Friendship. *American Philosophical Quarterly* 24 (4): 349-356.

Arendt, Hannah. 1975. Remembering W.H. Auden, *New Yorker*, January 20, 1975

Aristotle. 2014. *Nichomachean Ethics*. Translated by Roger Crisp. Cambridge: Cambridge University Press.

Auchincloss, Louis. 1962. Edith Wharton and her New Yorks. In Irving Howe (ed.), *Edith Wharton: A Collection of Critical Essays*. Englewood Cliffs, NJ: Prentice-Hall.

Austen, Jane. 2015 [1816]. *Emma*. New York: Penguin.

Badhwar, Neera Kapur. 1987. Friends as Ends in Themselves. *Philosophy & Phenomenological Research*, 48: 1–23.

Bailey, Olivia. 2023. Empathy, Sensibility, and the Novelist's Imagination. In Patrik Engisch & Julia Langkau (eds.), *The Philosophy of Fiction: Imagination and Cognition*. 218-239. New York: Routledge.

Blum, Lawrence. 2003. Personal Relationships. In R. G. Frey & Christopher Heath Wellman (eds.), *Blackwell Companion to Applied Ethics*. Oxford: Blackwell.

Booth, Wayne C. 1988. *The Company We Keep: An Ethics of Fiction*. Berkeley, CA: University of California Press.

Carroll, Noël. 1996. Moderate Moralism. *British Journal of Aesthetics* 36 (3): 223-238.

Carroll, Noël. 1998. Art, Narrative, and Moral Understanding. In Jerrold Levinson (ed.), *Aesthetics and Ethics: Essays at the Intersection*. 126-60. Cambridge: Cambridge University Press.

Cholbi, Michael. 2017. Grief's Rationality, Backward and Forward. *Philosophy and Phenomenological Research* 94 (2): 255-272.

Claparède, Edouard. 1928. Feelings and Emotions. In M.L. Reymett (ed.), *Feelings and Emotions: The Wittenberg Symposium.* 124-139. Worcester, MA: Clark University Press.

Cocking, Dean and Jeanette Kennett. 1998. Friendship and the Self. *Ethics* 108 (3):502-527.

Conee, Earl. 1994. Phenomenal Knowledge. *Australasian Journal of Philosophy* 72 (2):136-150.

Conrad, Joseph. 2012 [1907]. *The Secret Agent*. London: Penguin.

Crane, Tim. 2019. The Knowledge Argument is an Argument about Knowledge. In Sam Coleman (ed.), *The Knowledge Argument*. 15-31. Cambridge: Cambridge University Press

D'Arms, Justin and Daniel Jacobson. 2000. The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophical and Phenomenological Research* 61 (1): 65-90.

Dancy, Jonathan. 2004. *Ethics Without Principles*. Oxford: Oxford University Press.

Delaney, Neil. 1996. Romantic Love and Loving Commitment: Articulating a Modern Ideal. *American Philosophical Quarterly* 33 (4): 339-356

De Sousa, Ronald. 1987. *The Rationality of Emotion*. Cambridge, MA: MIT Press.

Deonna, Julien A. 2006. Emotion, Perception and Perspective. *Dialectica* 60 (1):29–46.

Deonna, Julien A. and Fabrice Teroni. 2012. *The Emotions: A Philosophical Introduction*. Routledge.

Deonna, Julien A. and Fabrice Teroni. 2015. Emotions as Attitudes. *Dialectica* 69 (3): 293-311.

Deonna, Julien A. and Fabrice Teroni. 2022. Emotions and Their Correctness Conditions: A Defense of Attitudinalism. *Erkenntnis*: 1-20.

Diamond, Cora. 2003. The Difficulty of Philosophy and the Difficulty of Reality. *Partial Answers: Journal of Literature and the History of Ideas* 2 (1): 1-26

Dokic, Jérôme and Stéphane Lemaire. 2013. Are Emotions Perceptions of Value? *Canadian Journal of Philosophy* 43 (2): 227-247.

Dokic, Jérôme and Stéphane Lemaire. 2015. Are Emotions Evaluative Modes? *Dialectica* 69 (3): 271-292.

Dorsch, Fabian. 2012. *The Unity of Imagining*. Berlin: De Gruyter.

Döring, Sabine A. 2007. Seeing What to Do: Affective Perception and Rational Motivation. *Dialectica* 61 (3): 363-394.

Feagin, Susan. 1996. *Reading with Feeling*. Ithaca, NY: Cornell University Press

Fodor, Jerry. 2008. *Lot 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fourie, Melike M., Henri G. L. Rauch, Barak E. Morgan, George F. R. Ellis, Esmè R. Jordaan and Kevin G. F. Thomas. 2011. Guilt and Pride Are Heartfelt, but not Equally so. *Psychophysiology* 48: 888-899.

Frankfurt, Harry G. 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy* 68 (1): 5-20.

Frankfurt, Harry G. 2004. *The Reasons of Love*. Princeton, NJ: Princeton University Press.

Friend, Stacie. 2012. Fiction as a Genre. *Proceedings of the Aristotelian Society* 112: 179-209.

Friend, Stacie. 2022. Emotion in Fiction: State of the Art. *British Journal of Aesthetics* 62 (2): 257-271.

Gaut, Berys. 1998. The Ethical Criticism of Art. In Jerrold Levinson (ed.), *Aesthetics and Ethics: Essays at the Intersection*. 182-203. Cambridge: Cambridge University Press.

Gibson, John. 2007. *Fiction and the Weave of Life*. Oxford: Oxford University Press.

Giustina, Anna. 2022. Introspective Knowledge by Acquaintance. *Synthese* 200 (2): 1-23.

Goldie, Peter. 2000. *The Emotions: A Philosophical Exploration*. Oxford: Oxford University Press.

Goldie, Peter. 2006. Wollheim on Emotion and Imagination. *Philosophical Studies* 127 (1): 1-17

Goldie, Peter. 2011a. Anti-empathy. In Amy Coplan & Peter Goldie (eds.), *Empathy: Philosophical and Psychological Perspectives.* 302-317. Oxford: Oxford University Press.

Goldie, Peter. 2011b. Intellectual Emotions and Religious Emotions. *Faith and Philosophy* 28 (1): 93-101.

Goldie, Peter. 2012. *The Mess Inside: Narrative, Emotion, and the Mind*. Oxford: Oxford University Press.

Goldman, Alvin. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goodman, Nelson. 1954, *Fact, Fiction and Forecast*, Cambridge, MA.: Harvard University Press.

Gordon, Robert M. 1986. Folk Psychology as Simulation. *Mind and Language* 1(2): 158-71.

Gregory, Dominic. 2016. Imagination and Mental Imagery. In Amy Kind (ed.), *The Routledge Handbook of Philosophy of Imagination*. 97-110. London: Routledge.

Hamilton, Christopher. 2003. Art and Moral Education. In Jose Luis Bermudez & Sebastian Gardner (eds.), *Art and Morality*. 37-55. London: Routledge.

Heal, Jane. 2003. *Mind, Reason and Imagination: Selected Essays in Philosophy of Mind and Language*. Cambridge: Cambridge University Press.

Helm, Bennett W. 2010. *Love, Friendship, and the Self: Intimacy, Identification, and the Social Nature of Persons*. Oxford: Oxford University Press.

Hume, David. 1960 [1739]. *A Treatise on Human Nature*. Oxford: Oxford University Press.

James, Henry. 2009 [1904]. *The Golden Bowl*. London: Penguin.

Jackson, Frank. 1986. What Mary Didn't Know. *Journal of Philosophy* 83 (5): 291-295.

James, William. 1884. What is an Emotion? *Mind* 9: 188.

Johnston, Mark. 1992. How to Speak of the Colors. *Philosophical Studies*, 68: 221–263.

Jollimore, Troy. 2011. *Love's Vision*. Princeton, NJ: Princeton University Press.

Kajtár, László. 2016. What Mary Didn't Read: On Literary Narratives and Knowledge. *Ratio* 29 (3): 327-343.

Keller, Simon. 2000. How Do I Love Thee? Let Me Count the Properties. *American Philosophical Quarterly* 37 (2): 163-173.

Kieran, Matthew. 2003a. Forbidden Knowledge: The Challenge of Immoralism. In Jose Luis Bermudez & Sebastian Gardner (eds.), *Art and Morality*. 57-73. London: Routledge.

Kieran, Matthew. 2003b. In Search of a Narrative. In Matthew Kieran and Dominic Lopes (eds.), *Imagination, Philosophy, and the Arts*. 69-87. London: Routledge.

Kolodny, Niko. 2003. Love as Valuing a Relationship. *Philosophical Review* 112 (2): 135-189.

Kolodny, Niko. 2010. Which Relationships Justify Partiality? General Considerations and Problem Cases. In Brian Feltham and John Cottingham (eds.), *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*. 169-193. Oxford: Oxford University Press.

Kraut, Robert. 1986. Love De Re. *Midwest Studies in Philosophy* 10 (1): 413-430.

Kundera, Milan. 1988 [1986]. *The art of the novel*. Translated by Linda Asher. London: Faber and Faber.

de Lafayette, Madame. 2020 [1678]. *La Princesse de Clèves, La Princesse de Montpensier et autres romans*. Paris: Gallimard.

LaFollette, Hugh. 1996, *Personal Relationships: Love, Identity, and Morality*, Cambridge, MA: Blackwell Press.

Lamarque, Peter and Stein Haugom Olsen. 1994. *Truth, Fiction, and Literature: A Philosophical Perspective*. Oxford: Oxford University Press.

Langkau, Julia. 2021. On Imagining Being Someone Else. In Amy Kind and Cristopher Badura (eds.), *Epistemic Uses of Imagination.* 260-278. New York: Routledge.

Lemaire, Stéphane. 2014. Norms for Emotions: Intrinsic or Extrinsic. In Julein Dutant, Davide Fassio and Anne Meylan (eds.), *Liber Amicorum Pascal Engel*. University of Geneva.

Lewis, Clive Staples. 2015 [1969]. *A grief observed*. London: Faber and Faber

Lewis, David K. 1990. What Experience Teaches. In William G. Lycan (ed.), *Mind and Cognition*. 29-57. Oxford: Blackwell.

Martin, Michael G. F. 2002. The Transparency of Experience. *Mind and Language* 17 (4): 376-425.

Matravers, Derek. 2014. *Fiction and Narrative*. Oxford: Oxford University Press.

Matravers, Derek. 2017. *Empathy*. Cambridge: Polity Press

McCarthy, Cormac. 2019 [2006]. *The Road*. London: Picador

McCrum, Robert. 2014. The 100 best novels: No 45 – The Age of Innocence by Edith Wharton (1920). *The Guardian*. 28 July 2014 https://www.theguardian.com/books/2014/jul/28/100-best-novels-age-of-innocence-edith-wharton-robert-mccrum

Meskin, Aaron. 2004. Aesthetic Testimony: What Can We Learn from Others about Beauty and Art? *Philosophy and Phenomenological Research* 69 (1):65–91.

Mikkonen, Jukka. 2021. *Philosophy, Literature and Understanding: On Reading and Cognition*. London: Bloomsbury Academic.

Molnar, George. 2003, *Powers: A Study in Metaphysics*, Oxford: Oxford University Press.

Mumford, Stephen. 1998, *Dispositions*, Oxford: Oxford University Press.

Müller, Jean Moritz. 2017. How (Not) to Think of Emotions as Evaluative Attitudes. *Dialectica* 71 (2): 281-308.

Naar, Hichem. 2013. A Dispositional Theory of Love. *Pacific Philosophical Quarterly* 94 (3): 342-357.

Naar, Hichem. 2018. Sentiments. In Hichem Naar & Fabrice Teroni (eds.), *The Ontology of Emotions*. Cambridge University Press.

Naar, Hichem. forthcoming. Love as a Disposition. In Christopher Grau & Aaron Smuts (eds.), *Oxford Handbook of the Philosophy of Love*. Oxford University Press.

Nehamas, Alexander. 2016. *On Friendship*. New York: Basic Books.

Nemirow, Laurence. 1980. Review of Nagel's *Mortal Questions*. *Philosophical Review* 89 (3):473-7.

Nemirow, Laurence. 1990. Physicalism and the Cognitive Role of Acquaintance. In William G. Lycan (ed.), *Mind and Cognition*. 490-498. Oxford: Blackwell.

Nietzsche, Friedrich, 2006 [1887]. *On the Genealogy of Morals.* Translated by Carol Diethe. Cambridge: Cambridge University Press

Nozick, Robert. 1989. Love's Bond. In *The Examined Life: Philosophical Meditations.* 68–86. New York: Simon & Schuster.

Nussbaum, Martha C. 1990. *Love's Knowledge: Essays on Philosophy and Literature*. Oxford: Oxford University Press.

Nussbaum, Martha C. 2001. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press.

Palmer, Frank. 1992. *Literature and Moral Understanding: A Philosophical Essay on Ethics, Aesthetics, Education, and Culture*. Oxford: Oxford University Press.

Peacocke, Christopher. 1985. Imagination, Experience, and Possibility. In John Foster and Howard Robinson (eds.), *Essays on Berkeley: A Tercentennial Celebration*. Oxford: Oxford University Press.

Perky, Cheves W. 1910. An Experimental Study of Imagination. *The American Journal of Psychology* 21 (3): 422 – 452.

Phillips, D. Z. 1972. Allegiance and Change in Morality: A Study in Contrasts. *Royal Institute of Philosophy Lectures*. 6: 47-64.

Prinz, Jesse. 2004. *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford: Oxford University Press.

Protasi, Sara. 2016. Loving People for Who They Are (Even When They Don't Love You Back). *European Journal of Philosophy* 24 (1): 214-234.

Ravenscroft, Ian. 1998. What Is It Like to Be Someone Else? Simulation and Empathy. *Ratio* 11 (2): 170-185.

Ripstein, Arthur. 1987. Explanation and Empathy. *Review of Metaphysics* 40/3: 465–482.

Roberts, Robert C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.

Robinson, Jenefer. 2005. *Deeper Than Reason: Emotion and its Role in Literature, Music, and Art*. Oxford: Oxford University Press.

Rossi, Mauro & Tappolet, Christine. 2019. What Kind of Evaluative States are Emotions? The Attitudinal Theory vs. the Perceptual Theory of Emotions. *Canadian Journal of Philosophy* 49 (4): 544-563.

Rorty, Amelie O. 1986. The Historicity of Psychological Attitudes: Love Is Not Love Which Alters Not When It Alteration Finds. *Midwest Studies in Philosophy* 10 (1): 399-412.

Sartre, Jean-Paul. 2004 [1940]. *The Imaginary: A Phenomenological Psychology of the Imagination*. London: Routledge.

Shiota, Michelle N., Samantha L. Neufeld, Wan H. Yeung, Stephanie E. Moser and Elaine F. Perea. 2011. Feeling Good: Autonomic Nervous System Responding in Five Positive Emotions. *Emotion* 11 (6): 1368-1378

Sibley, Frank. 2001. *Approach to Aesthetics: Collected Papers on Philosophical Aesthetics*. Oxford: Oxford University Press.

Solomon, Robert C. 1993 [1976]. *The Passions*. Indianapolis, IN: Hacket.

Stern, Tom. 2015. Against Nietzsche's 'Theory' of the Drives. *Journal of the American Philosophical Association* 1 (1):121-140.

Steward, Helen. Unpublished Manuscript. Demeanour.

Stocker, Michael. 1976. The Schizophrenia of Modern Ethical Theories. *Journal of Philosophy* 73 (14): 453-466.

Stocker, Michael. 1981. Values and Purposes: The Limits of Teleology and the Ends of Friendship. *Journal of Philosophy* 78 (12): 747-765.

Tanner, Michael. 2003. Ethics and Aesthetics are -. In Jose Luis Bermudez & Sebastian Gardner (eds.), *Art and Morality*. 19-36. London: Routledge.

Tappolet, Christine. 2016. *Emotions, Values, and Agency*. Oxford: Oxford University Press.

Taylor, Gabriele. 1976. Love. *Proceedings of the Aristotelian Society* 76: 147-164.

Telfer, Elizabeth. 1971. Friendship. *Proceedings of the Aristotelian Society* 71: 223-241.

Teroni, Fabrice. 2007. Emotions and formal objects. *Dialectica* 61 (3): 395-415.

Teroni, Fabrice. 2016. Emotions, Me, Myself and I. *International Journal of Philosophical Studies* 24 (4): 433-451.

Teroni, Fabrice. 2017. The Phenomenology of Memory. In S. Bernecker and K. Michaelian (eds.), *Oxford Handbook of Philosophy of Memory*. 21-33. Oxford: Oxford University Press.

Thomas, Laurence. 1987. Friendship. *Synthese* 72 (2): 217-236.

Tye, Michael. 2008. *Consciousness Revisited: Materialism Without Phenomenal Concepts*. Cambridge, MA: MIT Press.

Tye, Michael. 2009. A New Look at the Speckled Hen. *Analysis* 69 (2): 258-263.

Väyrynen, Pekka. 2006. Moral Generalism: Enjoy in Moderation. *Ethics* 116 (4): 707-741.

Vendrell-Ferran, Íngrid. 2022. Imagine What It Feels Like. In Anja Berninger & Ingrid Vendrell Ferran (eds.), *Philosophical Perspectives on Memory and Imagination*. 251-271. London: Routledge.

Walsh, Dorothy. 1969. *Literature and Knowledge*. Middletown, CT: Wesleyan University Press.

Walton, Kendall L. 1990. *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.

Walton, Kendall L. 2015. *In Other Shoes: Music, Metaphor, Empathy, Existence*. Oxford: Oxford University Press.

Waugh, Evelyn. 1978 [1945]. *Brideshead Revisited*. London: Eyre Methuen.

Wharton, Edith. 1999 [1920]. *The Age of Innocence*. Ware: Wordsworth.

Williams, Bernard. 1976. Persons, Character, and Morality. In *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. London: Fontana.

Wilson, Catherine. 1983. Literature and Knowledge. *Philosophy* 58 (226): 489-496.

Wolf, Susan. 1982. Moral Saints. *Journal of Philosophy* 79 (8): 419-439.

Wollheim, Richard. 2015 [1980]. *Art and Its Objects*. Cambridge: Cambridge University Press.

Wollheim, Richard. 1984. *The Thread of Life*. Cambridge: Cambridge University Press.

Zagzebski, Linda Trinkaus. 2017. *Exemplarist Moral Theory*. Oxford: Oxford University Press.

Zahavi, Dan. 2008. Simulation, Projection and Empathy. *Consciousness and Cognition* 17 (2): 514-522.

Zahavi, Dan. 2010. Empathy, Embodiment and Interpersonal Understanding: From Lipps to Schutz. *Inquiry: An Interdisciplinary Journal of Philosophy* 53 (3): 285-306.

Zahavi, Dan and Philippe Rochat. 2015. Empathy≠Sharing: Perspectives from Phenomenology and Developmental Psychology. *Consciousness and Cognition* 36: 543-553.