9-2015

Problems, Puzzles, and Paradoxes for a Moral Psychology of Fiction

Katherine Tullmann
Graduate Center, City University of New York

Problems, Puzzles, & Paradoxes for a Moral Psychology of Fiction

by

Katherine Tullmann

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2015

The dissertation has been read and accepted for the Graduate Faculty in Philosophy to satisfy the dissertation requirement for the degree of Doctor of Philosophy.

Noël Carroll

_____ _____

Date Chair of the Examining Committee

John Greenwood

_____ _____

Date Executive Officer

Jesse Prinz (adviser)

Noël Carroll

David Rosenthal

_____

Supervisory committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Problems, Puzzles, and Paradoxes for a Moral Psychology of Fiction

By Katherine Tullmann

Adviser: Dr. Jesse Prinz

The goal of my dissertation is to provide a comprehensive account of our psychological engagements with fiction. While many aestheticians have written on issues concerning art and ethics, only a few have addressed the ways in which works of fiction offer problems for general accounts of morality, let alone how we go about making moral judgments about fictions in the first place. My dissertation fills that gap. I argue that the first challenge in explaining our interactions with fiction arises from functional and inferential arguments that entail that our mental states about fictional entities are non-genuine. This means that our mental states during our engagements with fiction are different in kind from typical beliefs, emotions, desires, etc. that we have in real-life contexts. I call this position the Distinct Attitude View (DAV). In its place, I propose a common-sense, standard attitude view (SAV): the idea that our psychological interactions with non-real entities can be explained in terms of the intentional content of those states as opposed to a distinct type of mental state. In expanding the SAV, I develop several independent accounts of social cognition, emotions, and moral judgments. I also show how the SAV can dissolve standard problems in the philosophy and psychology of aesthetic experience: the paradox of fiction, the problem of imaginative existence, and the sympathy for the devil phenomenon, amongst others.

Acknowledgements

Many thanks to the following people who edited and commented on various chapters of this dissertation: my advisor, Jesse Prinz; my committee, Noël Carroll and David Rosenthal; my readers, John Greenwood and Barbara Montero. I also thank Berit Brogaard, Stephanie Ross, Nickolas Pappas, Ryan Dechant, Zoe Jenkin, Josh Keton, Jamie Lindsay, Rachel McKinney, David Neely, Jake Quilty-Dunn, and Denise Vigani for their comments. I am incredible grateful for their support as I worked on this dissertation. Many thanks also to my parents, Dennis and Karen Tullmann, for their continued and loving support.

Contents

Chapter 1: The Distinct Attitude View

1. Stranger than Fiction

You wake up early one morning and turn on the television. A newscaster informs you of a pair of brothers, known on the streets as "the Saints," who have committed multiple murders in the name of God. You haven't heard of the Saints and you are intrigued. The newscaster describes how the brothers only kill "bad people": Russian mobsters, murderers, hitmen, and the like. You then watch as the newscaster interviews people on the streets of your city. Their reactions to the Saints vary widely, from open approval of their actions ("Where can I sign up?", "Love the Saints, man! I think they're doin' a great job!"; "They're making the world a safer place") to revulsion ("Who are they to be judge and jury?"; "Killing for *good*?!"; "This is going to create something so much worse…") to fear ("I'm afraid to leave my house at night!"), and even social awareness ("*You*'re giving them the power!"; "You'll walk into a kid's bedroom, it's gonna' be there—Batman, Superman, and the Saints").

You, too, develop an opinion of the Saints as the broadcast continues. Maybe you think that their vigilante crime-fighting is good for society. After all, it's not like they are killing innocent people. Or maybe you are appalled by the hypocrisy of religious fanatics "playing God." Or perhaps you just worry about even more crime in the city. The thought of a mob war or underground street fight fills you with apprehension and anxiety for your own safety.

Then the screen cuts to credits. Slightly abashed, you realize that the news story was just that—a story. As the credits role, you discover that the newscast was a scene from a film called *The Boondock Saints*, a story about two Bostonian, Irish-Catholic brothers who decide to take the law into their own hands after the murder of a close friend.

Your interpretation of the Saints broadcast has changed. You now know that the events portrayed are a part of a fictional story. Before you believed that the Saints and their actions were real; now you know that they are not. You realize that the people and future consequences that you were afraid of are not actual threats. But what about your *appraisal* of the Saints? Has that changed as well? Say that you decided that the potential cost of the Saints' vigilante justice outweighs the benefits. While watching the clip, you pronounced a moral judgment that their actions were wrong. Now, after the clip is over, are you forced to accept that this moral judgment was not a *real* judgment?

There are some philosophers (Walton 1990, Currie 1995, Currie and Ravenscroft 2002, amongst others) who would answer this question with a resounding "yes." Our moral judgments of fictions—as well as our beliefs, desires, and emotions—are not *genuine* judgments. Rather, our evaluation of the Saints is a pretend, imaginary, or judgment-*like* mental state. This claim leads to a peculiar, yet surprisingly popular, interpretation of the previous example. When we watch a fictional film, read a novel, or see a play, the beliefs, desires, judgments, and emotions that we experience are not standard or stereotypical beliefs, desires, judgments, or emotions. Yet, when we consider a similar—even an almost *exactly* similar—event on BBC or in *The New York Times*, our mental states are stereotypical simply because their object is part of the actual world.

I will call this the Distinct Attitude View (DAV; based on Schroeder and Matheson's Distinct Cognitive Attitude View, or DCAV. See Schroeder and Matheson 2006). The DAV holds

a dominant position in both the philosophical and psychological work in aesthetics. As we will see, some versions of the DAV claim that our mental states towards fictions are not ordinary states because they arise from functionally and inferentially different input and result in distinct cognitive and behavioral outputs from analogous real-life scenarios. We can contrast this view with a standard attitude view, or SAV, which states that we utilize the same types of mental states during our engagements with fictions and real-life; differences in behaviors and responses to fictions can be explained in terms of the content of standard beliefs, emotions, judgments, etc. I will defend a version of the SAV in this dissertation.

Proponents of the DAV typically focus on the nature of our emotions or beliefs about fictions. I am interested in our *moral* evaluations of fiction, like our judgment of the Boondock Saints. While many philosophers and psychologists have accepted some form of the DAV, few have considered its implications for moral issues and moral psychology—Gregory Currie's 1995 paper, "The Moral Psychology of Fiction" and Elisabeth Camp's unpublished work "Perspectives in Imaginative Engagement with Fiction" are exceptions. Indeed, my project here is largely a critical response to these papers. Furthermore, many aestheticians have written on issues concerning art and ethics, but only a few (e.g. Currie 1995, Nussbaum 2001) have addressed the ways in which works of fiction offer problems for general accounts of morality, let alone how we go about making moral judgments about fictions in the first place. Finally, moral psychologists ask questions concerning our moral judgments of *real-life* people and events. Although they utilize fictional cases and thought experiments in their studies, few have discussed how the ontological status of these cases may affect the participants' responses. Ironically, the way in which we know about real-life/standard/typical mental states is through psychological studies that often utilize fictions in their experiments.

This dissertation attempts to fill these gaps. In doing so, I offer a comprehensive moral psychology of fiction. As we will see, determining whether our mental states towards fictions are stereotypical or not will be paramount to our understanding how we judge fictional characters, how we act on those judgments, and the propriety of those judgments. The questions I will address in the following chapters are very similar to those found in papers on "real-life" moral psychology. For instance: do moral judgments involve conscious reasoning? What role do emotions play in our moral judgments? Can we make unconscious moral judgments? Moral psychologists also ask questions concerning how people *acquire* moral values: can we use our reason to come to an evaluation? Are moral values culturally-constructed or innate? Finally, are we always motivated to act on our moral evaluations of a person or situation? I will spend the majority of the following chapters evaluating potential responses to these questions. I argue that fictions often pose unique problems for moral psychology, but that these problems can be avoided if we base our moral theories on adequate accounts of emotions, moral judgments, and social cognition.

In the present chapter, I will analyze the implications of the DAV in greater detail. Along the way, I will highlight the explananda of a moral psychology of fiction, the basic features of our interactions with fictions that any adequate account should address. In §2, I discuss the "puzzles of fiction," a set of intriguing problems that have been taken to suggest that there is an asymmetry between how we respond to real life and analogous fictional scenarios. §3 illustrates two versions of the DAV in greater detail: a theory of *make-believe* and theory of *imagination*. §4 evaluates three arguments that are typically used to support the DAV: arguments from the functional and inferential roles of mental states and another from research in cognitive neuroscience. In §5, I raise four further challenges to the DAV. I conclude by providing a sketch of my own view— the standard attitude view, or SAV—and a look ahead to the following chapters.

2. The puzzles of fiction

Some of the most persuasive evidence for the DAV stems from the asymmetries between our responses to analogous real-life and fictional scenarios. Some of these asymmetries are captured by the following puzzles of fiction. Since Plato's *Republic*, philosophers and psychologists have taken these puzzles to present problems for theories of our mental attitudes towards fiction. One of my goals in this dissertation is to diffuse these puzzles; I will show why they are not really so puzzling to begin with. In this section, I will give a brief overview of the different puzzles that we will encounter in the chapters to come.

First, the *paradox of fiction* concerns whether we may have genuine emotions towards fictional entities and states of affairs. It is largely understood that emotions play a significant role in an organism's survival by tracking features of our environment that may impact our well-being (see Damasio 1994; LeDoux 1996; Currie 1995; Prinz 2004a, etc.). Yet, this point seems difficult to reconcile with the claim that we experience genuine emotional responses towards fictional objects. Perhaps we can broaden our notion of "well-being" in terms of an object's general impact on our mental or physical state. But even this seems problematic for fictional objects, for we know that what happens in a fiction generally has little impact on us at all—and, even if it *does* impact us in some way, we still need an explanation of why and how. There are three different issues that must be untangled here: the question of whether fictional objects have an impact on our well-being, the nature of our emotions about fictions, and, finally, the scope of the behavioral and cognitive

dimensions of our emotional responses to fictions. Any adequate solution to the paradox of fiction should address all three. I will discuss the paradox of fiction more thoroughly in chapter 5.

There are several further puzzles that arise primarily in respect to our moral judgments of fiction: the *puzzle of disparate response*, the *sympathy for the devil phenomenon* (Carroll 2004 and Camp, *in preparation* for terms), the *puzzle of imaginative resistance* (Walton 1994, 2006; Gendler 2000, 2006; Currie 2002; Moran 1994), the *puzzle of moral motivation* (indirectly addressed in Walton, 1990), and the *question of moral learning*. The puzzle of disparate response covers the asymmetrical mental responses that we have to fictions and real-life people. This includes a character's moral indiscretions. The most obvious examples of this can be found in fictions that portray violence and law-breaking in a positive light, such as mafia and heist films like *The Godfather*, *Ocean's Eleven*, and *Le Cercle Rouge*. Westerns like *The Searchers* or *The Good, the Bad, and the Ugly* romanticize violence and portray "heroes" with whom we probably would not want to be friends in our real lives. Even superheroes and vigilantes would enrage many audiences if they were discovered in their own cities (an issue interestingly explored in films like *The Boondock Saints*, *Watchmen, and The Dark Knight*).

The sympathy for the devil phenomenon is a special instance of the puzzle of disparate response. This puzzle concerns why we have positive responses (identification, empathy, sympathy, concern, etc.) to anti-heroes and villains in a fiction, characters that we would abhor in the actual world. Both historical and contemporary fictions abound with examples of alluring anti-heroes, from Milton's charismatic Satan in *Paradise Lost*, to Nabokov's Humbert Humbert, and *Breaking Bad*'s Walter White. In each case, we cheer on and have positive emotional and moral responses to these characters whose real-life counterparts we would find repugnant. I will address the sympathy for the devil phenomenon in chapter 7.

The puzzle of imaginative resistance describes a similar phenomenon. Fictions often present us with highly improbable or metaphysically impossible scenarios (maybe even *logically* impossible ones), such as time travel, the ability to create and enter another person's dream, and the existence of vampires. Most audiences take these impossibilities in stride. We accept that time travel, dream invasion, and vampires are possible in the fictional world. Yet there seem to be some fiction/real world inconsistencies that an audience member will not (and maybe *should* not) accept. These include moral actions, and values that are positively portrayed in a fiction, but we would not endorse in real life (see, e.g. Currie 2002; Gendler 2000 & 2006; Hume, 1757/1994; Moran, 1994; Walton 1994 & 2006; Weatherson, 2004). Tamar Gendler describes the puzzle of imaginative resistance as: "the puzzle of explaining our comparative difficulty in imagining fictional worlds that we take to be morally deviant" (2000, p. 56). I will return to this puzzle in chapter 8.

Another moral puzzle I will consider has received only tangential treatment in the literature, but will nevertheless play a considerable role in the chapters to come. We tend to think that our moral judgments are *motivating*; if we believe that a certain action is morally wrong, then we generally will not perform it and we may take measures to prevent the action from occurring. Yet we are not motivated to act on our moral judgments about fictions. This is the problem of moral motivation. If morals are necessarily (or even only *usually*) motivational, why do we rarely act upon our moral judgments of fictional characters and situations? For example, witnessing horrific gang violence in a film rarely leads one to call the police. The easy answer to this is that we know that fictions are not real, so we know that we cannot call the police in a story to report a fictional crime. Another, more pressing question is whether the recognition of a morally bad action in a fiction should somehow cross the ontological boundary into real-world judgments, and

influence subsequent real-world behavior. Part of this puzzle rests on issues concerning the functional role of our mental attitudes towards fictions; I will address this in §4 of this chapter, and again in chapter 6.

Finally, I will consider the possibility of learning about morality from fictions. It may seem strange that we can learn anything about the actual world from works of fiction. After all, fictions generally do not purport to tell the truth. Yet many people would argue that fictions are an important resource for learning about other perspectives and cultures, even if the characters and places are not real. It also seems like we can develop our knowledge of folk psychology and morality by the scenarios fictions present, and we can hone our emotional and moral abilities by "practicing" on fictions. These positive accounts of learning from fiction need some explanation of how we can learn about the real world from fictional material. I will consider several positive accounts of learning from fictions (Lamarque 2009, Nussbaum 1995, Robinson 2005) as well as several arguments against this idea (Currie 2013) in the final chapter of this dissertation.

The puzzles of fiction comprise a serious challenge to a moral psychology of fiction. Any adequate theory of our psychological interactions with fiction must account for them. Indeed, the puzzles of fiction are often taken to be so problematic that philosophers have concluded that we need to reinterpret how we understand our mental attitudes about fictions and even propose new types of mental attitudes. I reject this move. For now, though, let's explore the adversary's position in more detail.

3. The Distinct Attitude View

Imagine that you are reading *The Hobbit*. Tolkien introduces you to a strange new world where,

amongst other magical occurrences, you fictionally encounter dragons. Your responses to Smaug are probably nothing like what they would be if you encountered the wily beast in real life; you would *fear* Smaug rather than feel curious or intrigued by him and you would run away from the creature rather than remain in your comfy chair. Indeed, we often do not physically react to the objects we see on a screen or read about in a novel. Our desires concerning the fate of a fictional character have a different relation to our other beliefs, thoughts, and desires than they would if that character was a real person. Call this *the asymmetry problem*: we behave and respond differently to fictional entities than we would to their real-life counterparts. We draw different cognitive inferences and make different emotional and moral judgments of them. What explains these differences?

3.1. Two interpretations

There are two general ways we can explain the asymmetry problem. One is to suggest that the *content* of our mental attitudes towards fiction is different than the content of our mental attitudes towards real-life objects. Call this the *content-based view*; this will be the basis for my SAV. The other account urges that we have different *types* of mental attitudes towards fictions and real-life events and objects. This is the *distinct attitude view* (DAV). In this section, I will present the general outline of both views and offer several reasons why we should prefer the former. In §4, I will expand upon, and reject, three typical arguments given in favor of the DAV.

The content-based view holds that our engagements with fictions and their real-life counterparts result in different behaviors and cognitive inferences due to a difference in the *intentional content* of our mental attitudes about each. In other words, our thoughts, beliefs, desires, emotions, etc. have different types of objects than they would in real-life situations. When

we engage with a fiction, we recognize that what we encounter is not real; the events in a fiction generally do not or did not actually happen. For example, when I see a production of *A Midsummer's Night Dream*, I recognize that the actors on the stage embody fictional characters that do not physically exist in our actual world. This recognition informs our further mental engagements with the fiction, both consciously and unconsciously.

The content-based view also requires that we tacitly employ a fictional operator when we *speak* about fictions. When I say "I believe that Demetrius treated Helena very poorly" I mean "I believe that Demetrius treated Helena very poorly [in the fiction]" (see Kripke 2011 and Thomasson 1999). The same holds for our mental attitudes; the fictional operator is implicit in our beliefs, thoughts, and desires about fictional entities (see also Matravers 1991 & 2014, Neill 1993). The recognition that the object of our mental attitudes does not exist is the backdrop against which they are formed. So I may have a belief "that Smaug is a wily beast [in the fictional world]" but not "Smaug is a wily beast" *tout court*.

There is quite a bit more to be said about the nature of our mental states and speech acts concerning fictions, as well as the metaphysical implications of the content-based view. I will expand upon some of these issues in chapter 3. For now, it is enough to note the differences between this type of view and the DAV. Importantly, the content-based view maintains the possibility that our mental attitudes towards fictions are not different in kind from those about actual objects. We treat characters as fictional objects, but we nevertheless have standard beliefs about them. The DAV, on the other hand, holds that a solution to the asymmetry problem requires a *distinctive cognitive attitude* that we employ in our engagements with fictions.

The nature of the distinct attitude varies by the theorist. Quite often, they posit a mental "box" or mechanism that we utilize exclusively when considering non-actual objects, including

fictional ones (see, for example, Nichols & Stich 2003), but also hypothetical and counterfactual thought, mental activities involving deliberation and decision-making, mental imagery, and mindreading (attributing mental states to other people).

Alternatively, we may use largely the same mental processes as in real-life situations, but these processes are run "off-line," disconnected from their typical functional and inferential output. The result is a different kind of mental state than we would have if we considered a real-life/actual object. The result is that we adopt *pretend* (Searle 1975; see also Kripke 2013), *simulated* (e.g. Walton 1996, Currie & Ravenscroft 2002), or *imaginary* beliefs, desires, emotions, and thoughts toward non-actual objects (e.g. Schroeder & Matheson 2006, Weinberg & Meskin 2006). These states are isomorphic to genuine mental attitudes and may even be phenomenologically indistinguishable from them. Still, we cannot say that they are the same type of state due to their distinct functional and inferential roles.

I want to resist this move. I will argue that a content-based view of our mental attitudes towards fiction can fully solve the asymmetry problem. This means that we do not need to posit a distinct cognitive attitude in the first place. Note that the content-based view does not entail that our experiences with fictions and real-life are exactly the same. Of course we respond differently to an actual, face-to-face snarling dog than we would to one that we encounter on a movie screen. The challenge will be to explain how, on my view, we can explain the behavioral differences if the mental states are of the same type.

The content-based view also does not imply that we have the exact same kinds of thoughts about fictional and real-life objects. A mental state's inferential role—what mentally causes it or what further mental processing it results in—could be roughly the same whether its object is non-actual or real. However, the mental state's intentional *content* differs in each case. This difference

may account for asymmetrical behaviors and cognitive inferences. The content-based view implies that our knowledge that fictional objects *are fictional* informs our mental attitudes towards it, but does not imply that the attitude is no longer of the same kind. It only implies that the attitude is *about* something different.

3.2. Make-believe

I will now present two versions of the DAV: a theory of make-believe and a theory of imagination. According to a theory of make-believe, our engagements with fiction draw on our capacities for imagination and pretense. While reading a novel, for example, we play a game of make-believe, creating a "fictional world" in which the propositions presented in the novel are fictionally true. So, in contrast to the content-based view, the mental state type during our engagements with fiction differs from real-life situations, but we treat the intentional content in each case the same—as if the object is a real one.

On this view, each reader is the participant and creator of her own fiction-based game. Fictional worlds are similar to the imaginary worlds that children create during their pretend play (Walton 1990; Currie 1990, Harris 2000). For our purposes, *pretense* is the activity of acting and thinking as if some proposition or state of affairs is true, while knowing that it is not. In their games of make-believe, a child may pretend that a couch is a house, the underneath of a dining room table is a dungeon, and a baseball bat is a magnificent sword—all while knowing that none of these are actually the case.

Theories of make-believe hold that something similar occurs when adults engage with fictions, but this time the props are largely imaginative. While reading the *Harry Potter* series, we

pretend that there is a world similar to our own in which witches and wizards go to a secret school called Hogwarts, and there's a Dark Lord lurking in the shadows. The novel itself acts as a prop, each line feeding into our fictional world and adding layer upon layer of detail to our game of make-believe. The game of make-believe extends to our psychological states: we pretend to believe that Harry Potter defeated the Dark Lord at the tender age of one and we pretend that it is true that Harry is the youngest Quidditch seeker in over one hundred years. Furthermore, we have fictional feelings about Harry and his crew and fictional desires for the young wizard to vanquish the Dark Lord.

A theory of make-believe attempts to explain two things: the ontology of fictions and the nature of our psychological states about them. Both rely on engaging in pretense-based games. Let's begin with the ontological question. Do fictions and fictional characters exist? What are fictional entities such that we can think about them, speak about them, etc.? Theories of make-believe have a ready response to these questions: fictional entities do not, strictly speaking, exist outside of one's game of make-believe. Instead, when we discuss fictional entities, we make merely pretend illocutions concerning pretend objects. Like an actor in a performance of *Hamlet*, we do not make genuine assertions when we speak about the goings-on of fictions. We merely pretend to do so as a part of the game (see also Searle 1975). Fictional entities are not objects that can be found in space or time, even if the images or words used as props can be (a painted figure, a film image of a person, or words that describe a villain).

Walton contrasts his view with a realist Meinongian theory, according to which fictional entities exist eternally as abstracta, similar to Platonic forms (1990; see also Sainsbury 2010 and Wolterstorff 1980). On this view, fiction entities are not created, but rather drawn upon and put together in creative ways by authors, filmmakers, dramatists, etc. Meinongian theories have come

under strong attack, and rightly so. However, contrasting a pretense theory with the Meinongian position ignores the more fine-grained concerning the existence of fictional entities. The middle ground includes other broadly realist theories, such the possible objects view (Lewis 1974) or the abstract artifact view favored by Kripke (2011 & 2013), Salmon (1998), Schiffer (1996), and Thomasson (1999).

Importantly, neither a realist nor pretense-based ontology entails that our mental states about fictions are distinct in type. It could be that we merely pretend that fictional objects exist when we think about and discuss them. Our mental states about those objects are the standard mental states, *about* something fictional. That is one possible way in which to understand the ontological and psychological questions of fiction: an anti-realist ontology of fictional entities coupled with a genuine attitude view of mental states. Unfortunately, this view raises a host of questions concerning the possibility of referring to nonexistent objects.

For this reason, many anti-realists about fictional entities do not go this route. Instead, they favor the DAV according to which our mental states towards fictional entities are not stereotypical mental states. One popular way in which to ground a psychology of make-believe is to adopt some form of *simulation theory* (ST). ST is typically used in cognitive science as a way to explain how we attribute mental states to others, especially to predict their behaviors and understand their emotional state (see Goldman 2006, Gordon, 1986, Currie & Ravenscroft 2002, Nichols et al 1996, Prinz 2002). Currie and Ravenscroft (2002) argue that imagination essentially involves the capacity to put ourselves in the place of another or our self in another place or time. Importantly, both Walton and Currie/Ravenscroft hold that the mental attitudes we adopt in our imaginings are substitutes for genuine ones; we have *imaginative* or *fictional* beliefs about fictions instead of

beliefs simpliciter. On their view, imagining simulates the role of other states, such as the role beliefs play in inferential processes (*ibid*, 49).

The upshot of a theory of make-believe—combining anti-realism about fictional entities and ST to explain our psychological states towards them—is that is allows proponents of the DAV a theoretically cohesive and elegant means in which to solve the puzzles of fiction. For example, we can solve the paradox of fiction by arguing that we do not possess ordinary types of beliefs about fictional entities. We feel "sympathy for the devil" because that emotion is distinct from our normal cognitive processing, which allows us to experience unique emotional and moral responses towards entities that do not match how we would respond to their real-life counterparts. And so on. However, if the theory of make-believe entails that our mental states towards fiction are non-stereotypical, then this proposal is not available to those who dismiss the DAV. The content-based view needs some alternative method to explain, or explain away, the puzzles of fiction.

3.3. Imagination

The proponent of a content-based view seems to be at a disadvantage in explaining another aspect of our engagements with fiction. Note that theories of make-believe rely heavily on our capacity to imagine things that are not the case. It is often assumed that imagination draws upon our powers of pretense and mental simulation. Another version of the DAV holds that imagination is necessary for our experiences with fiction; fictions call us to imagine certain propositions and states of affairs; this may include distinct, imaginary kinds of mental states. Since imagination seems to be involved in both the creation and enjoyment of fictions, then, on this logic, pretense or simulation must be as well. But if imagination requires pretense, and pretense requires distinct attitudes, how

can the content-based view explain this extremely important facet of fictions?

It is worth asking whether our imaginative capacities do, in fact, require pretense. I think that this is doubtful. Some imaginative engagements do not require that we hold some non-actual proposition to be true. We may imaginatively entertain a proposition, consider an alternative state of affairs, or "call to mind" an object without automatically holding that it is the case. Entertaining, considering, and calling to mind may not require pretending that the given state of affairs is true. Perhaps this is the case with fiction; we consider or entertain fictional propositions but do not accept their truth in the actual world.

The act of imagination is also sometimes thought to require simulation. On this view, when we imagine that something is the case, we run our mental states off-line, disconnected from their typical functional and inferential roles. We simulate typical mental processes. 'Imagining that X is true' means that we simulate that X is true. Again, I think that it is a mistake to equate imagination and simulation. Perhaps imagining that something is the case sometimes involves simulation. But simulation is not required for imagining in general. We can consider or entertain non-actual states of affairs without simulating mental processes as if that they are actual.

If I am right here, then we must take care to distinguish between imagination, pretense, and simulation. Unfortunately, we are now left with the question of what imagining actually amounts to, if not one of these other capacities.

Neil van Leeuwen (2013) helpfully distinguishes between three types of imagining, each of which may (but arguably need not) be involved in our experience with fictions. The first type is *constructive imagining*. This kind of imagining involves developing the content of mental representations, such as objects, ideas, premises, etc. (*ibid*, 221). We create the objects of our mental states. This kind of imagining could be involved when an author writes a novel, or even

when a scholar pens an article. But constructive imagining could also be involved on the recipient's (the person who reads the novel or the article) side. We often do much more than casually entertain thoughts concerning fictional characters. We consider implications for a character's future or we may consider what the story would be like if some event had or had not occurred. For example, we may wonder whether the boy will ever get the girl or if the vampire has a heart after all. We may also imagine what it would be like to be Elizabeth Bennett when she first arrives at Pemberley or Bruce Wayne when he learns the truth about Bane's past.

Van Leeuwen's second type of imagining is *attitude* imagining. Here, we regard the content of our imagination as fictional rather than as describing the world as it really is. I take it that this is the most important type of imagining that needs to be discussed. This type of imagining is also sometimes called *propositional* imagining; we are imagining *that* something is the case.

Lastly, *imagistic* imagining is when we visually, auditorily, or in some other way use mental imagery to represent some intentional content. We can close our eyes and imagine the opening measures of Beethoven's *Eroica* or picture a character from our favorite novel enact her most daring exploits. This type of imagining is not greatly at stake in my discussion. I will, however, briefly discuss this type of imagining in §4 since it is sometimes used as evidence in favor of the DAV.

Despite its widespread popularity in theories of fiction, I reject the claim that imagining—and especially attitude imagining—is necessary to explain our interactions with fictions. This is a hard line to sell, but one that I think is acceptable if we keep the distinctions I've made in this section in mind. I deny that constructive imagining and imagistic imagining are necessary for our engagements with fiction, although the former may be required for the construction of fictions (writing a novel, creating a film, etc.). And although we may often have a great deal of mental

imagery involved in our experiences with fiction, we do not need to. It may not be a part of our conscious experience of fictions that we imagistically imagine what Jane Eyre looks like, for example.

Some philosophers argue that some kind of imagining is required in order for us to make sense of a fictional story (Camp, unpublished manuscript). Imagining in this sense means "holding in mind objects or states of affairs that are not present to us." This is roughly equitable to *considering*, *supposing*, or *entertaining* a proposition. My claims in this dissertation are compatible with the imagination thesis if this is all that is meant by imagining. However, those who support the DAV typically take imagining to mean more than merely considering or entertaining a proposition–they make claims concerning the nature of the mental states or processing involved, often appealing to imaginary, offline mental states. I object to *this* aspect of the imagining thesis. Finally, it is important to distinguish this kind of imagining (if it *is* imagining) is different from both pretense, make-believe, and simulation, all of which make stronger claims concerning the types of mental activity involved in these processes.

I do not deny that these further kinds of imagination *can* play a role in how we engage with fictions. Rather, I argue that imaginative engagement is not *necessary* for our engagements with fiction (see also Matravers 2014). I will take a neutral stance on the ontology of fictional characters, although in general I favor an abstract artifact view. I think that my SAV is compatible with most ontological theories of fictional entities. In contrast, the proponent of the DAV generally holds that questions concerning the nature of fictional entities go hand in hand with how we think and feel about them. Non-actual objects need to be understood in terms of atypical mental states. I will return to this in chapter 3. For now, though, let's consider several arguments in favor of the DAV.

4. Three arguments for the DAV

The main theoretical support for the DAV stems from a folk psychological and functionalist understandings of the nature of mental attitudes. Beliefs, desires, judgments, etc. have characteristic behavioral and inferential roles, understood in terms of inputs from stimuli and their cognitive and behavioral outputs. Real-life beliefs and their imaginative counterparts might utilize much of the same causal and inferential pathways to bring about certain responses, but they employ significantly different inputs and result in very different output (see, e.g. Currie 1990, Currie & Ravenscroft 2002, Weinberg & Meskin 2006, Schroeder & Matheson 2006). Furthermore, research in psychology and cognitive neuroscience seems to support the idea that engaging with non-actual objects utilizes *many,* but not *all*, of the same neural pathways as our mental activities concerning actual things (Kosslyn 1997; Kosslyn, Thompson & Ganis 2006). These studies help to explain why our reactions to imaginative activity are often so robust, but may also suggest that we utilize a distinct attitude in our engagements with fiction (Damasio, 1994; Schroeder & Matheson 2006).

In this section, I will present and critique three arguments in favor of the DAV: the argument from functional role, the argument from inferential role, and the argument from neuroscience. Doing so clears the path for my own content-based view by eliminating the main motivation behind positing a distinct cognitive attitude to begin with.

4.1. The argument from functional role

Proponents of the DAV argue that mental states such as beliefs and desires can be identified in terms of their characteristic functional roles. Consider beliefs. It is generally understood that our beliefs about real-life and non-actual objects utilize many of the same causal and inferential pathways. However, they have significantly different inputs and result in very different outputs. Our real-life, everyday beliefs are about things that we can see, touch, and hear— things that exist spatio-temporally (concretely). These objects act as the input for our everyday, stereotypical beliefs. In contrast, beliefs about fiction (and other non-actual objects) do not take actual, concrete objects for their input. Rather, they are about *fictions*, however that is cashed out ontologically. The output of our mental processing about actual and fictional objects is also different; they result in different kinds of behaviors. So while our beliefs about real life and our beliefs about fictions are *similar* in many ways, proponents of the DAV hold that they are different enough to constitute a distinct kind of mental state. The result is that we have belief-*like* attitudes towards the content of fictions, but not beliefs simpliciter.

A folk-psychological account of mental attitudes helps to motivate this claim. Standard belief-desire psychology holds that a belief that *X*, together with a desire to *Y*, motivates action, all things being equal. This position is echoed in the literature on fiction. Proponents of the DAV take it as a basic fact about our cognitive processing that beliefs generally motivate action (see Walton, 1990; Currie & Ravenscroft, 2002; Schroeder & Matheson, 2006; Weinberg & Meskin, 2006). If there is no resulting motivation to act, then we do not have a stereotypical belief. We will see that the same idea applies to other mental states, including emotions. As Walton argues: "Fear emasculated by subtracting its distinctive motivational force is not fear at all" (1990, 202).

This statement captures the neo-behaviorist, or "purely motivational" conception of belief.

David Velleman describes this view as follows: "all that's necessary for an attitude to qualify as a belief is that it disposes the subject to behave in certain ways that would promote the satisfaction of his desires if its content were true" (Velleman 2000). The problem with this view, as Velleman notes, is that it cannot account for a variety of cases that we want to describe as instances of belief. For example, there may be cases in which a subject "believes that *P*, but this belief does not bring with it a *disposition* to act in *P*-concordant ways because of some feature of the subject" (Gendler 2008, 653; emphasis added). Another possibility is that the subject believes that *P*, but does not *act* in *P*-concordant ways because the belief lacks behavioral implications altogether. Finally, there are cases in which a subject is disposed to act in *P*-concordant ways, but does act on them because she lacks *other* required beliefs (*ibid*, 653).

Each of these three cases may apply to our interactions with fictions. Perhaps I am not disposed to act on a belief about a fiction because I lack the relevant desire. Or maybe my belief about a fiction simply lacks any behavioral implications. My belief that Sherlock Holmes lives at 221B Baker Street does not necessarily motivate me to *do* anything. Finally, I could lack other relevant beliefs for action, such as the belief that Sherlock Holmes is a flesh-and-blood person.

Of course, there are different ways in which we can understand the nature of belief besides the neo-behaviorist conception (see Tullmann & Buckwalter, *forthcoming*). For instance, L. Jonathan Cohen argues that the belief that *x* is *the disposition to feel that x is true* (Cohen 1989). This notion of belief could work for our engagements with fictions if we take the truth condition to be *true according to the fiction*, as we saw in the previous section. For now, let's assume that the functionalist picture of belief is correct. Does it necessarily entail the DAV? I will argue that it doesn't; we can also evaluate the motivational role of belief in terms of the state's *content*.

Consider Currie and Ravenscroft's (2002) argument for imaginary mental states. They

maintain that beliefs serve as part of an inferential network, motivating not only action, but also other beliefs, desires, and thoughts. The authors posit that beliefs about a fiction are run "off-line," disconnected from their normal behavioral and cognitive networks. The same goes for other activities with non-actual content, such as hypothetical thought and pretend play. These states are not stereotypical beliefs, but rather *imaginative* ones, because stereotypical beliefs are understood in terms of the behavior that they produce.

If Currie and Ravenscroft are correct, then we do seem to have a distinct attitude for fictional and other activities. Before we accept this conclusion, however, I want to explore two other possible interpretations of the functional data that do *not* entail the DAV. First, note that the main charge against a content-based view is that we do not have the right kind of behavioral responses to fictions; our "beliefs" about them do not play the *right kind* of functional role for motivating action. In contrast, I argue that we *do* have genuine beliefs about our imaginative activities, but we do not act because due to the influence of other facts and judgments that we have (that the object is not real, for one)—even if we are motivated to on some subconscious or precognitive level.

Second, our beliefs about non-actual objects do sometimes result in actual behaviors. This is especially clear in cases of hypothetical deliberation and pretend play. For example, a child pretending to have a tea party will go through the motions as if she was really having a tea party; someone deliberating on an important decision may go through some of the actions she would need to perform in order to test a potential outcome, as if she were really going through with that outcome. I will make the case for acting on the basis of beliefs about fictions as well.

If I'm right, then we are at least sometimes motivated to act on our beleifs about fictions, even though we know that the object of our belief is not real. Let's begin with my second claim:

we sometimes act in response to fictions. There are two ways in which fictions might motivate actions the first relates a fictional scenario to an analogous real-life situation; the second has to do with judgments of the fiction *qua* fiction. According to the first kind of motivation, there may be cases in which engaging with a fiction and imaginatively considering it will shape our future actions in our daily lives. For example, after watching *Blood Diamond*, I make sure that the jewelry I purchase comes from legitimate sources that do not utilize slave labor. It may be quite common for fictions to shape our lives in this way; we "export" information from the fiction and apply it to real-life. In this respect, beliefs about fictions can play a role in our everyday behavioral networks.

An obvious response to this account of fiction-based action is that it does not really address the relevant type of behavioral motivation. We need some account of our behavioral responses to fictions *qua* fiction. It may seem like it is extremely rare that we are ever motivated to act in this way. For example, we are seldom motivated to attempt to speak to or walk alongside a fictional character. I am typically not motivated to punish a fictional villain for his crimes that take place in the fiction, and will not leave a theater demanding that justice be served (if it is not in the narrative). But there are other ways in which we may actually act towards things that we know are not real. This is especially true if we consider acting out of an *emotion*. I cry for Jane Eyre when I find out about Mr. Rochester's wife; I would cry for her too if she were my actual best friend. When I watch a scary film I clutching my armrests, cover my face with my hands, I yell at protagonist to "watch out!" These are all types of actions—and they are all performed even though I know that the character, monster, etc. do not actually exist. Of course, we do not act in the exact same way that we would if that fictional monster were in our town. But we cannot say that beliefs and other mental states about fictional objects do not yield *any* functional output. Nor can we say that the functional output is always distinct in kind. At least some of our actions-out-of-emotion towards

fictional entities will match those we would perform towards real-life people.

I take it that the proponent of the DAV will not be wholly satisfied with this (admittedly) limited response. So let's consider my first claim: we are often motivated to act on the basis of our beliefs about fictions, even if those actions are not carried to fruition. On this account, it is possible that we are motivated to act in response to our interactions with fictions even if we do not follow through with the action. Perhaps my belief that there is a monster oozing towards the screen (think Walton's Green Slime) may dispose me to act. Again, some of these behaviors are carried out—I gasp and worry my hands in anxiety. Other behaviors are not; I do not run screaming towards the exit, yelling for the other characters to go hide. Why does my belief that there is a green slime terrorizing the town motivate me to act in some ways but not others?

I hold that some motivated actions do not come to fruition because they are inconsistent with my recognition that the slime is not real. My belief that the green slime is going to harm me may have *unconsciously primed* me to act; my motor system may have been activated and prepared for flight. In the neo-behaviorists' terms, I am *disposed* to act. But that action is ultimately blocked from occurring due to other beliefs and thoughts that I may have—most importantly, the disbelief in the slime monster's actual existence.

Whether or not unconscious, automatic behavioral priming does, in fact, occur is a question for empirical investigation. We can imagine a study in which we compare a participant's behavioral reactions to beliefs about fictions to analogous beliefs and actual, carried-out behaviors; if the relevant neural circuitry matches, even in part, we would have some evidence that we are primed for action based on our beliefs about fictions. Indeed, there is some evidence that this is the case (see, for example, Freedberg & Gallese 2007).

The psychologist Paul Harris (2000) provides further evidence for this view, particularly

in terms of our emotional responses to imaginings. Very young children tend to be overcome by the emotions caused by their imaginative activities—e.g. fearing the monster under the bed or witches that they saw in a movie—even if they know that the fiction is not real. Older children and adults are generally able to regulate and override these emotions, but still may often become absorbed in their imaginative activity. We get "lost" in a film or novel and experience strong emotional reactions, emotions that color our real-life activities.

One way to explain these reactions is to consider the evaluative process of emotions. Emotions are inherently value-laden insofar as they represent objects in our environment that may affect our well-being. Harris points out that we "do not appraise the inputs from a perspective outside of that imagined framework" (*ibid*, 66). Rather, we appraise an imaginative activity from *within* the imaginative framework itself. In that respect, we may be at least motivated to act, even if it is on an unconscious or subpersonal level. Typically we do *not* act, though, because other regulatory processes kick in (what Harris calls "control processes"). For example, we remind ourselves that the scary scenario is just a fiction or focus on how the story was produced instead of its scary content. These are ways of detaching ourselves from the fiction. Doing so overrides the standard functioning of our ordinary emotional responses.

I think that something similar is going on in a great deal of our interactions with fictions. We may be have low level motivations to act on but generally do not because of other regulatory thoughts, judgments, and beliefs. The motivation to act may not be able to override these other factors, especially the knowledge that what we encounter does not actually exist. So our mental states towards fictions still have their characteristic functional role, but it is blocked by some other attitude. Our dispositions to act sometimes do not result in action. Having the belief that fictional objects are not real may act as a kind of traffic light for behavior, stopping some behaviors from

ever coming to fruition while allowing others to proceed. This does not mean that the beliefs involved in actually motivating those behaviors are any different in kind from ordinary, real-life beliefs. It does imply that there is some deterring factor (namely, some other mental state) that allows for some behavior while barring others.

4.2. The inferential role argument

Even if I am right that the functional argument for the DAV fails, one could still claim that there is a difference in the *inferential* role of mental states about actual and non-actual objects. This difference is enough to claim that mental attitudes about fictions are different in kind from those about actual things. A mental state's inferential role includes both its *meaning* and *relation to* other mental states. For example, on conceptual-role semantics, for a state to possess the content that it does is simply for it to play a certain role in cognition. A belief that *X* will play a role in bringing about further thoughts, beliefs, desires, etc. The belief is also caused by other mental states. Importantly, the difference between the inferential role of a stereotypical belief about an actual object, and a belief-like state about a fiction, may be a difference that *makes a difference*. We do not have stereotypical beliefs concerning fictions because our real-life and fictional beliefs play different types of inferential roles. They are caused by, and result in, different types of states and so possess different meanings. I will show that this conclusion is unfounded.

To begin, one could argue that the difference in inferential role results from a difference in intentional *content*. If that is the case, then we have an explanation for why the real-life and fictional beliefs seem to play different inferential roles: they are beliefs about different things! Consider the following example. When you watch the film *American Psycho*, you witness Patrick

Bateman's dalliances from one exorbitantly-priced dinner to another, examining and trading business cards, discussing the merits of Versace versus Gucci suits, and picking up prostitutes to torture and kill. You form certain beliefs about Patrick: "Patrick is a superficial maniac"; "Patrick is a product of our materialistic culture"; "Patrick is a moral monster." These beliefs were caused by other beliefs, thoughts, emotions, etc. and in turn cause other beliefs, thoughts, and emotions.

Now imagine that you encounter someone just like Patrick in your real life: a superficial, charming, handsome Wall Street exec that you discover has a secret passion for torture and murder. Your beliefs about this Wall Street exec play roughly the same inferential role as those in the fiction because they have roughly the same content. We believe that this Patrick is a superficial maniac, etc., and this causes further beliefs, thoughts, and judgments. However, no matter how similar this real Wall Street exec is to Patrick, he is not exactly the same. Specifically, our belief about the fictional Bateman has a fictional operator; we believe "that Patrick Bateman is a Wall Street exec [in the fictional world of *American Psycho*]." This difference could make the difference in terms of the state's inferential role: different mental content, different belief, different inferential role.

It could be argued that it's possible for a fictional scenario to have the exact same content as a real-life one. The proponent of the DAV could argue that even if the content were *exactly* the same, our beliefs about fictions and a real-life counterpart would play distinct inferential roles.

We can vary the *American Psycho* example slightly to highlight this point. The fictional story of Patrick Bateman is the same, but this time imagine that, unbeknownst to Bret Easton Ellis (the author of *American Psycho*) there is, in fact, a real Patrick Bateman who is exactly like the Patrick of the story. The proponent of the DAV would have to show that in this case our mental states still play distinct inferential roles, despite an exact similarity in content. But it is not clear

how one could make this argument. In this case, it is just an accident that the Patrick Bateman of the fiction is exactly the same as the Patrick Bateman of the actual world. A reader would still treat the fictional character *as* a fictional character because she does not know any different. Even if the reader finds out about the real Patrick, she may still have different types of mental states about them—namely, that the first one is Patrick of the story, and the other is Patrick of the actual world.

In fact, I think that it is generally impossible for a fictional and real-life scenario produce the exact same mental content. If they did, it is not clear what would separate the fiction from the non-fiction—what would make the fiction *fictional*? Wouldn't we just call *American Psycho* a *non*-fiction? Now, one could respond that it is just a coincidence that Ellis wrote a story that exactly imitates the life of a real person. That does not make his work any less a fiction (Borges' "Pierre Menard, Author of the *Quixote*" comes to mind here). Moreover, we might speak of the Napoleon of Tolstoy's *War and Peace* as if he is the actual Napoleon, for example, and so equating the two in our mind. But this is a mistake. Here we have a fictional portrayal of a (once) actual person, but differs in important ways from the actual Napoleon. Even in this case, as with the two Patrick's, we are left with a difference in intentional content, the difference of being in a fictional story versus the actual world. This results in two different mental states. We do not need to bring in a distinct mental attitude to explain a difference in inferential role.

There is a further response we can make against the DAV, one similar to my second claim concerning the functional role argument. Just because the mental state has different behavioral or cognitive output does not mean that it is a distinct type of state or that there are distinct cognitive mechanisms involved in their processing. The distinct inferential role may be taken as evidence for the DAV, but it does not entail it. Regardless of the fictional/real-life mental attitude's content, it will still play some inferential role. We can explain this role in the same way we would a belief

with actual content, in terms of the mental attitudes it causes. The resulting difference between the real-life and fictional intentional roles can be explained by the other beliefs, desires, thoughts, and emotions that brought it about. This will generally include the belief that a particular character or scenario is not real or only exists in the fictional world. This places the belief in a distinct role, but does not seem to require that we adopt a whole new cognitive mechanism.

I conclude that the argument from inferential role does not entail the distinct attitude view. A difference in content likely explains the inferential differences in our mental attitudes towards fictions. Moreover, we do not need to posit a distinct cognitive attitude even if the content is exactly the same.

4.3. The argument from neuroscience

The functional and inferential role arguments are the most common justifications for the DAV. Some DAV proponents, however, also believe that research in cognitive neuroscience supports their view. Here I will briefly consider some of this research and how it has been utilized by various philosophers, in particular, Timothy Schroeder and Carl Matheson (2006). My response to their arguments is quite similar to my response to the previous two: there are multiple interpretations of the neuroscientific data, and these interpretations are either just as plausible as the distinct attitude interpretation, or more so.

First of all, we must clarify what the neuroscientific research is best suited to explain. It seems clear that much of the neuroscientific research on imagining concerns van Leeuwen's third category, imagistic imagining. For example, Stephen Kosslyn argues that about two-thirds of the processes utilized in vision are also employed in visual imagery. The similarity arises because

mental imagery relies on previously organized and stored information; the difference arises because some of the low level processes involved in *actual* perception do not occur in imagistic "seeing," "hearing," etc. (Kosslyn 1997). Alvin Goldman (2006) argues that there is also evidence that we use many of the same processes when we *imagine* a face as when we actually see it (O'Craven and Kanwisher, 2000). The fusiform gyrus is activated in both cases; lesions to the fusiform face area impair both face recognition and the ability to imagine faces (Damasio et al. 1990). Finally, Goldman argues that there are neural similarities in motor execution and motor imagination. Imagining a moving hand utilizes the same neural mechanisms that would normally be used when we actually move that hand (Parsons et al. 1998, Freedberg & Gallese 2006).

This research provides evidence for the similarities and differences between imagistic imagining and actual perception or movement. I doubt that anyone would deny that there is a difference between mental imagery and actual perception—and, furthermore, that at least *some* different cognitive and neural mechanisms are involved in each. But this is not the kind of imagining that is at stake in the distinct attitude debate. Rather, the significant question seems to be whether, in spite of their similarities to real-life situations, we need to posit a distinct mental state to explain our *attitude* and *constructive* imaginings about fictional content. Again, based on the previous two arguments, it is unclear that we do.

Nevertheless, Schroeder and Matheson build on the neuroscientific imagistic data in order to account of our emotional reactions towards fictions in a way that supports the DAV. Their arguments are similar to Goldman's; research in cognitive neuroscience shows that we employ mental attitudes that are "akin to beliefs [etc.] in structure and in some of their effects, but distinguished from beliefs in others" (2006, 20). Their task is to clarify these similarities and differences.

To begin, the authors discuss how sensory and quasi-sensory representations are produced. The stimulation of the sensory organs produces neural signals that then produce patterns of activity in the brain known as *unimodal sensory representations* (*ibid*, 26). These unimodal representations are thought to come together to form multimodal representations that can be evoked through multiple sensory systems. Multimodal representations portray things like snakes, trees, and other objects in our environment that can be experienced by the use of more than one sensory modality. Our semantic memory contains dispositions to token representations of these objects. These are activated by sensory stimulation just as unimodal representations are (*ibid*, 26).

Importantly, multimodal representations are signaled to a number of different neural targets, including the orbitofrontal cortex, the affective division of the striatum and the amygdala (*ibid*, 26; see Rolls 2000, Mello & Villares 1997 and LeDoux 1996). Each of these areas is involved in producing affective responses and emotional feelings. The orbitofrontal cortex is largely involved in the "discrimination of rewarding and punishing stimuli, in order to influence feelings, visceral responses, and decision-making" (*ibid*, 27). The affective division of the striatum is involved in producing reward signals that affect emotional feelings. Finally, research by the neuroscientist Joseph LeDoux suggests that the amygdala is involved in basic emotional responses to stimuli, resulting in bodily reactions such as the increased heart rate, sweating, breathing, and facial expressions characteristic of fear and other emotions (*ibid*, 27; see also LeDoux 1996).

Schroeder and Matheson argue that not only are these affective areas stimulated in *real-life* situations, they are also activated in our *imaginative* activities, including engagements with fictions. In fact, there are no distinct anatomical mechanisms involved in our emotional responses to actual or non-actual objects (*ibid*, 28). For instance, our uni-and-multimodal representations are activated when we perceive something scary in a film. This information is sent to the affective

regions of the brain just as if we encountered something frightening in our actual environment. The same applies to what the authors call "free-floating imaginative stimuli" not directly caused by external stimuli (*ibid,* 29). There is some neuroscientific evidence to support this claim (see Kosslyn et al, 1993). For example, research suggests that a monkey's representational systems respond powerfully to puppet faces and even "extremely schematic two-dots-over-a-line-faces" (Schroeder & Matheson 2006).

The authors take this evidence as powerful support for the DAV. I agree that it certainly establishes the first half of their claim: our emotional responses to fictions are akin to those about actual objects. But they have not yet established the second, crucial part of their claim that our emotional responses to fictional and actual objects are inherently different kinds of states. What is their argument for the distinct attitude? It should look familiar. The authors argue that "although fictional and otherwise imaginary stimuli have many of the same effects as 'real' stimuli do, they obviously do not have all the same effects, or else people would leap onto stages in order to prevent murders, and so on" (*ibid*, 29). They continue:

> Perhaps, then, coming to have an imaginative faculty as a [distinct cognitive attitude] comes down to learning to treat representations that one creates oneself, or representations created by what one grasps to be fictions or the like, as not warranting the sorts of behavioral responses that they would were the representations created externally by non-imaginary events and objects: learning, that is, to give such representations a distinct functional role (*ibid*, 34).

So rather than explain the necessity of a distinct cognitive attitude in terms of data from cognitive neuroscience, Schroeder and Matheson fall back on the same kind of functional differences we have already considered. This is essentially the same argument that we discussed above and it is subject to the same kind of critique. The functionalist argument does not entail the DAV. The

differences can be explained purely in terms of the intentional content of the mental state. Alternatively, we could be disposed or motivated to act towards fictions, but other judgments, beliefs, and regulatory processes prevent us from doing so.

Interestingly, these explanations are compatible with a further point that Schroeder and Matheson make concerning the influence of background knowledge and dispositions on our responses to imaginative activity. It may be that these factors, including top-down judgments about the ontological status of fictions, keep our behaviors in check even if we are subpersonally and automatically primed or motivated to act or respond emotionally to them. This is consistent with the neuroscientific data the authors present. Engagements with fiction (and other non-actual objects) can generate actual emotional responses because they employ the same neural mechanisms as real-life situations do. It is also quite possible that the uni-and-multi-modal representations involved in emotional responses do not "pay attention," as it were, to the existence of the thing that they represent before signaling the relevant brain areas. Recognizing that the object of our interaction is not real requires top-down judgments from different cognitive pathways (see Damasio 1994, Harris, 2000 & LeDoux, 1996). The emotion goes through.

This section constitutes my preliminary remarks against the DAV, clearing the way for my own standard attitude view. I will return to these arguments in the following chapters.

5. Resisting the DAV

The DAV could provide us with a unified explanation of our psychological interactions with fictions, accounting for all of the puzzles of fiction in terms of one general approach. Furthermore,

work on the imagination and simulation lends the view empirical credibility and a potential integration with broader psychological theories.

I have already rejected the three main arguments in support of the DAV: arguments concerning the functional role, inferential role, and neural processing of our mental states (particularly beliefs) do not entail a distinct cognitive attitude. What I have not shown is why we should prefer a content-based, or standard attitude view, over a distinct attitude view. I will spend the next four chapters attempting to show why the SAV is preferable to the DAV. Here are several preliminary suggestions.

First, one concern with the DAV is that many accounts build their theory on a flawed notion of the nature of the mental states in question (particularly emotions). If we get things right on in terms of the ontology and psychology of mental states (or at least put forward the best view we can) then it is unlikely that a distinct attitude is warranted to begin with.

The second critique is related to the first. Those who argue in favor of a distinct attitude rarely explain *what*, exactly, the distinct attitude amounts to. As we've seen, theorists generally try to explain the distinctness of our imaginative attitudes in terms of functional role (behavioral outputs) or inferential role (mental outputs); our mental states towards fictions do not lead to the kinds of thoughts and behaviors that they would for real-life objects. This problematically assumes a straightforward functionalist view of mental states that, while attractive in terms of folk psychology, does not accurately capture the nature of how mental states motivate action or inferential processes. I will argue against this assumption in my dissertation. Doing so undermines the motivation behind the DAV to begin with; if mental states are not individuated in terms of functional role, then we may in fact have standard mental states towards fictions, despite the fact that we do not act towards fictional objects as we do towards real-life ones.

Third, one can argue from a principle of parsimony that there is no need to posit a distinct mental attitude if typical ones have the same explanatory power. I will attempt to show that we simply do not need to posit a DAV if we adopt an appropriate view of emotions and moral judgments, as well as suitable explanations for the puzzles of fiction.

Finally, the DAV (especially one utilizing a theory of make-believe or simulation theory) does not seem to be able to account for our actual phenomenological—that is, conscious, possibly introspectible—experiences with fictions. Our emotions, beliefs, desires, and other mental states towards fictions feel natural and relatively automatic, not like we are playing a game of make-believe, or simulating a possible course of action. In fact, our mental states about fictions often do not seem any different than those about actual things.

The proponent of the DAV may dismiss the phenomenological worry by arguing that the game of make-believe or simulation takes place unconsciously and, after some practice, quite rapidly. Some of the imaginings involved in a game of make-believe are often deliberate and consist in conscious, occurrent mental states. But others are spontaneous, unconscious, and automatic. We do not tell ourselves to begin imagining what is going to happen to our favorite television character. We simply do it, sometimes without realizing it. Walton says that when this happens our imaginings "have a life of our own" and we feel less like an author than a spectator to the imagining (Walton 1990, 14). I grant that this may be true for simulation (an issue I'll return to in chapter 4). But it is hard to imagine that we engage in an unconscious game of make-believe—how would the game work? What would the rules be, and how would it be initiated? Walton and other theorists of make-believe claim to have a theoretical answer these questions. What they do *not* seem to be able to explain is the phenomenology of our actual experiences.

6. The Standard Attitude View

Based on my discussion so far, we can isolate four questions that an adequate moral psychology of fiction must address:

1. Why do we often have asymmetrical psychological responses to analogous fictions and real life situations? Call this *the asymmetry problem*.
2. What is the nature of our mental attitudes toward fictions, especially those involved in moral judgments? Call this *the fictional attitudes problem*.
3. How, and to what extent, are we motivated to act in response to fictions, especially morally? Call this, as we have seen, *the problem of fictional motivation.*
4. What psychological theory can adequately account for our actual phenomenological experiences toward fictions, particularly in terms of our moral judgments? Call this *the experiential problem*.

The issues that I have raised against the DAV in the preceding sections should at least *suggest* that a different approach to a psychology of fiction is in order. I will argue for a standard attitude view (SAV), which, as the name suggests, holds that we form standard mental states concerning fictions. This implies that we can understand our mental states about fictions in much the same way we would any actual context—a simple enough claim, but one that has been strongly challenged by proponents of the DAV. Arguing for this position requires that I explore the alternate

interpretations to the functional, inferential, and neuroscientific data typically employed to support the DAV that I offered in the previous section.

It is a tall order for any theory to answer each of the above four questions. Nevertheless, in what follows I will argue that my SAV can explain them just as well or better than the competing DAV theories. Some of the different versions of the DAV and SAV may be able to address several of these questions, but I will show that only my view can adequately account for them all.

I will develop my own moral psychology from the "ground-up." To begin, I will discuss one further potential worry against the SAV: *the illusion theses*, the idea that we are under either a cognitive or perceptual illusion while engaged with fictions. However, developing my moral psychology of fiction does not only require that I debunk the DAV. I must also develop a theory that can explain our actual moral experiences with fictions. In chapter 3, I introduce a key concept for doing so: *the fictional stance*. I will then sketch a picture of my own theory of how we understand the mental states of fictional entities in chapter 4 in terms of a modified theory-theory for mindreading, while also critiquing other potential theories. In chapter 5, I will address the unique problems that emotions pose for a psychology of fiction. This will involve a discussion of the paradox of fiction, the nature of emotions, and the intelligibility of these responses. In so doing, I will develop my multi-level appraisal theory of emotions, which will serve as the backbone of my theory of moral judgments about fictions. I will take up that issue in Chapter 6 in which I argue for a sentimentalist, multi-level appraisal theory of moral judgments. I will also summarize my conclusions from the previous chapters and weaving together the various strands into my SAV.

The remaining three chapters will each discuss a different puzzle of fiction: the puzzle of disparate response/sympathy for the devil phenomenon in chapter 7, the puzzle of imagination resistance in chapter 8, and the question of moral learning in chapter 9. I will also return to the

four questions from this section and show how I have resolved them with my SAV.

While my arguments predominantly apply to our interactions with fictions, they also carry over into other areas of philosophy and cognitive science. Indeed, our mental states towards fictions are often important counterexamples to general theories of emotions, perception, and mindreading. So it is important to take them into account when we develop our general psychological and philosophical theories. But there are three other ways in which the underlying question of this dissertation—the DAV versus the SAV—is philosophically significant. First, there is an independent question of whether we should explain aspects of human behavior in terms of the content of our mental states or in terms of new mental attitudes. If both content and attitude-based views have a great deal of explanatory power, then how should we adjudicate between them? I take my work here to cast a mark in favor of the content-based approach.

Second, individuating and defining mental states is a complicated and controversial area of philosophy of mind. Fictions provide us with interesting, complex, and challenging examples and counterexamples for theories of mental state individuation. Finally, this dissertation covers very broad range of philosophical topics, from the nature of perception and belief to moral learning and emotional rationality. Indeed, I will sketch several positive theories in the following chapters: the modified theory-theory, an appraisal theory of emotions *and* moral judgments, and a theory of emotional responsibility. Indeed, this dissertation is as much an exploration of the nature of our mental states as it is an examination of our experiences with fictions.

Chapter 2: The fictional illusion theses

1. The "magic" of fiction

This chapter focuses on a potentially perplexing aspect of our interactions with visual representations, and especially visual fictions (including film, paintings, pictures, even video games): it seems like, in some cases, that visual fictions can play tricks on us. We may "suspend our disbelief" while engaged with fictions so that we come to believe that fictional objects are real and events actually occur.

The notion of suspending disbelief was first introduced by the British Romantic poet Samuel Coleridge, who argued that we suspend our disbelief in the nonexistence of fictional objects during our engagement with fictional stories (Coleridge 1817). This supposedly explains our emotional responses to fictional entities; we emotionally respond to them because we believe that they actually exist!

While it's certainly true that we sometimes become very *absorbed* in fictions, it is a genuine question whether we actually forget or are tricked into believing that fictional characters and events actually exist. For example, watching Christopher Nolan's 2006 film *The Prestige*, we may come to believe that the two magicians—Robert Angier and Alfred Borden—are real people, that their magical rivalry actually occurred, and maybe even that *their* illusions are real. Generally speaking, we know that the objects in fictional film are nonexistent, just like we know that magician's "magic" isn't real. The question is whether we can we be tricked, perhaps momentarily, into

believing otherwise.

Following Noël Carroll (Carroll 2008), let's characterize this challenge as *the illusion thesis*: we fall prey to some kind of illusion during our engagements with fiction. So far, we have been discussing one type of illusion, concerning belief. Carroll characterizes *two* types of potential illusion:

1) The Cognitive Illusion Thesis (CIT): one might come to believe that fictional events, characters actually occur or exist. This would mean that when I watch *The Prestige*, I come to believe that Angier and Borden are real and the events in the film actually take place.

2) The Perceptual Illusion Thesis (PIT): we are committed to the existence of the represented fictional objects on a perceptual level. This implies that we perceive fictional entities like Angier in the same way that we do real-life objects and, further, that our perception is somehow implies the existence of the perceptual object.

Different artistic media may commit us to one or both of these illusions. For example, reading a novel might subject one to a cognitive illusion, but not a perceptual illusion. Perceptual illusions generally apply to visual fictions, although we can imagine someone listening to an audiobook in her car falling prey to the perceptual illusion that the narrator is an actual person.

Many philosophers reject both versions of the illusion thesis out of hand. Indeed, what is the motivation behind them? One potential argument in favor of the illusion thesis comes from our dispositions to act on the basis of our beliefs or perceptions of fictional objects. In chapter 1, I argued that we may be motivated to act while we are engaged with a work of fiction. However, many of our motivated behaviors will never come to fruition because of competing judgments concerning the fictional status of the work. This argument formed the basis of my attack against the DAV (distinct attitude view), in favor of the SAV (standard attitude view). Our beliefs about fictions are ordinary beliefs that play typical functional roles, even if those beliefs do not

necessarily lead to action.

Interestingly, one could argue that *any* motivation to act in response to fictions, even if that motivation is entirely unconscious and never amounts to actual behavior, implies that we believe in the existence of the fictional object. This would suggest that my SAV position is committed to the CIT, and perhaps the PIT as well. Doesn't *any* kind of emotional response imply that we believe in the existence of fictional entities, either on a perceptual or cognitive level? Of course, it's worth pointing out the functionalist assumption behind such claims, as I noted in the last chapter. If we assume strong version of functionalism about mental states, such that beliefs and percepts can be individuated and identified primarily in terms of functional role, then our low level behavioral and emotional responses may suggest that we believe that a zombie (for example) is real, or that we perceive it as such.

I want to resist the illusion theses, partly by resisting their functionalist assumption. I face the challenge of developing an alternative explanation for these subconscious motivations to act, an explanation that does not appeal to an illusory belief in the existence of fictional entities. Furthermore, one may be tempted to think that illusory attitudes do not count as standard beliefs, desires, and emotions, or that they are unstable grounds for further genuine attitudes about fictions. Such views could open the door to a pretense theory of fiction; a suspension of disbelief in the nonexistence of fictional entities could imply that the viewer is engaged in some kind of pretend play or game of make-believe; we pretend that fictional objects are real. This would be problematic for the SAV.

I grant that we may be victims of *some* illusions while engaged with fictions. Most prominently, we might be under the illusion that objects actually move while we watch a film due to the speed of the projection (see Carroll 2008). We might also be under the illusion that a

particular actor has changed appearances (with makeup or prosthetics) or that we witness a magic trick (as in *The Prestige*). I do not deny that we are often "victims" of these kinds of illusions. Such illusions are commonplace, and generally easy to explain in terms of the material and construction of fictions. It is also possible that *sometimes* we are under more general illusions while engaging with fictions such that we perceive or believe that the fiction is real. Perhaps we encounter a performance of a play that takes place in the streets of our city and, being unfamiliar with the play, we do not realize that those are actors.[1] This would be a genuine case of a cognitive illusion.

In this chapter, I will argue that both the illusion theses are false if we think of them as general accounts of our experiences with fiction. It is not a part of our general experience with fictions that we believe or perceive fictional objects as existing. To make this argument, I will begin by drawing out further implications of the illusion theses and provide some preliminary remarks against them, beginning with the CIT. I will argue that we do not fall prey to the CIT, but that standard remarks made against the thesis are inadequate. In fact, there are three versions of the CIT; standard responses against the CIT can only account for two of them. The third version of the CIT is, I argue, actually a form of the PIT. This means that the CIT boils down to a question about perception and perceptual content. I will explore the PIT in §3 and §4. §5 addresses the possibility of the cognitive penetration of perception and how it may bear on the illusion theses. I will conclude by briefly presenting an alternative explanation of the behavioral and emotional responses that served as motivation for the illusion theses, and explanation that is compatible with the SAV.

[1] Many thanks to John Greenwood for pressing me on this point.

2. The cognitive illusion thesis

Carroll presents and critiques both illusion thesis in his book, *The Philosophy of Motion Pictures*. He characterizes the CIT as follows: during the duration of a movie, the viewer believes that the objects she encounters are real, so she treats them as she would any real object. As we've seen, the CIT essentially constitutes a suspension of disbelief in the objects and events of the fiction (Coleridge 1817, Hurka 2001). For example, a viewer suspends her disbelief in the nonexistence of the magician Angier while she watches *The Prestige*. Doing so allows her to recognize fictional objects as real things. This leads to some low level (subpersonal, automatic, non-deliberate) behavioral and emotional responses towards fictional entities.

What is the nature of the belief implied by the CIT? This question is not typically addressed in the literature on this topic. The assumption is that we're discussing *inferential beliefs*, which are traditionally 'cognitive' in nature; they result from inferential processes working together with current content, information about the world, and stored background knowledge (see Lyons 2005). This includes beliefs like "Moscow is the capitol of Russia." Some beliefs, however, arise directly from perceptual capacities. We may believe that the sky is grey (for example) by merely perceiving a grey sky, without further inferential processing. These are *perceptual beliefs*. I will show that both types of belief maybe subject to the CIT.

The CIT could potentially explain a wide-range of our experiences with fictions, including how we recognize objects in fiction as well as our cognitive interactions with them (emotions, thoughts, desires, etc.). However, most philosophers seem to think that it would be patently absurd to accept the CIT, since it seems to entail that we would act towards a character in just the same way that we would act towards a real person. As Katherine Thomson-Jones points out:

> I am able to appreciate the vivid depiction of an army of zombies surging forward with arms outstretched, the use of special effects or highly emotive music, the importance of the scene for the narrative, and so on. Surely, if I had suspended my belief that the zombies are fictional, I would be too frightened to appreciate film in this way (2008, p. 107).

The problem is that the majority of our behaviors towards fictions (or lack thereof) are inconsistent with the idea that we even temporarily suspend our disbelief about the reality of a fiction. We do not act as if we believe that the fiction is real. The same idea works for other mental attitudes, such as desires and emotions. Moreover, our conscious experience of watching a film is also antithetical to the cognitive illusion thesis. If asked, we would deny that fictional entities are real. We would also deny that we were tricked into believing otherwise.

These remarks comprise the standard responses to the CIT: we do not act as if we believe that fictional objects are real and we do not have the conscious experience of believing that they exist. The standard response provides adequate evidence against the CIT much of the time, especially if we accept the functionalist idea that beliefs have certain functional and inferential roles that often lead to behavior; if the relevant behavior is missing, then the belief probably is too.

However, there is a way to interpret the CIT that makes it much more of a challenge than writers typically acknowledge. In fact, I think that there are three ways to cash out the CIT:

> CIT#1). We have a conscious belief that the objects of fiction actually exist (spatio-temporally). E.g. I have a *conscious* belief that Hamlet exists.
>
> CIT#2). We have an unconscious belief that the objects of fiction actually exist (spatio-temporally). E.g. I *unconsciously* believe that Hamlet exists.
>
> CIT#3). We have an unconscious perceptual belief that the objects of fiction actually exist (spatio-temporally). E.g. I have a *perceptual belief* that Hamlet exists.

We now have to ask ourselves whether the standard responses to the CIT still work in light of this reinterpretation. Only CIT#1 entails that we consciously believe in the concrete existence of

fictional entities. A film viewer would actually believe, while watching a zombie film, that those very zombies are out there somewhere. Now, if this is true, then surely the standard responses would be correct—we would run screaming from the movie theater. We do not, so CIT#1 fails.

The same basic argument applies to CIT#2. It is generally accepted that both conscious *and* unconscious beliefs can influence and motivate behavior (see Rosenthal 2008). Indeed, our motivations for action are often unconscious; we are not always consciously aware of what feature of our thoughts or environment causes us to speak or act. For example, consider what unconscious beliefs are in play when you choose an apple at the grocery store, type a paper on your laptop, or ride your bike to work. This implies that unconscious thoughts, judgments, and beliefs can cause behavior. If CIT#2 is right, then an unconscious belief in the existence of an object of fiction may often motivate behavior. For example, we may have an unconscious belief that the zombies on a screen are real, and so we may be motivated to run screaming from the theater. The fact that we do not run screaming from the theater suggests that we have neither a conscious nor unconscious belief in the existence of the zombies—if we grant the functionalist assumption. So CIT#2 also fails.

What about CIT#3? Do the standard responses count against it? Unfortunately, they do not. CIT#3 rests on the idea that we can have perceptual beliefs that may contradict our consciously held inferential beliefs. So, for example, I may consciously (cognitively) disbelieve that Hamlet exists, while at the same time I have a perceptual belief that he *does* (see Quilty-Dunn, forthcoming, for a defense of CIT#3). I watch a zombie film and perceive the images on the screen as zombies. It could be that the processes involved in seeing the image and judging it to represent a zombie involves a belief that we are in the presence of an actual zombie. Perhaps recognizing an image of a certain object automatically commits us to a perceptual belief that we are in the presence

of that object. We have a *perceptual belief* that the zombie exists. Recall that perceptual beliefs are formulated directly from perceptual evidence and not from inferential cognitive processes. For example, I may directly formulate the belief that it is raining outside from looking out of my window; I do not need to draw any further inferences to believe this.

This means that our perception of the zombie-images causes us to have the perceptual belief in their existence. And perceptual beliefs may in turn motivate some actions and emotional responses. This interpretation of the CIT is compatible with my suggestion in chapter 1 that we may be primed, disposed, or motivated to act on the basis of fiction even if these actions do not result or manifest in actual behaviors. After all, if the functional role is fulfilled, aren't we experiencing *some* kind of belief, one that motivates our low level responses towards the fictional entity? For instance, it could be that perceptual beliefs explain how one's motor cortex primes for action while face to face with real people, but also when faced with paintings and statues of people (see Freedberg & Gallese 2007, Gallagher 2013; see Goldman 2006 for an overview). Perhaps we form an unconscious perceptual belief that the person in a painting or statue is real. This belief does not manifest in actual behavior (running from the theater) since it is blocked by other beliefs and judgments. But the illusion still occurs.

I will explore an alternative explanation of the low level behavioral and emotional data in the last section. For now, it's worth noting how we have reinterpreted the CIT. The standard responses worked against the notion of global cognitive illusion, but, at the same time, the CIT was clearly false. My reinterpretation of the illusion theses is more charitable towards pro-illusionists and more philosophically and scientifically interesting. The CIT boils down to a question about *perception,* about whether we perceptually experience fictional entities as existing before us. I will now examine this claim.

3. The perceptual illusion thesis

The PIT states that we have the perceptual experience of seeing real things while engaged with a visual fiction. We perceive the material of which the fiction is constructed in a non-illusory way—e.g., the image on our television or computer screen, or the paint on a canvas. We also perceptually recognize representations on a screen or canvas as being of particular objects: people, trees, animals, places, etc. The question is whether we perceive these objects in an illusory way. Carroll compares our visual experience of motion pictures to traditional perceptual illusions such as the Müller-Lyer illusion or perceiving a straw in a glass of water as bent when it is really straight (Carroll 2008). In each case, our visual system leads us to perceive something in the world as being some other way than it actually is. In the case of fiction, we would perceive the objects represented as real objects.

There are two criteria required to get the PIT off the ground. First, our visual experience of the illusory object must be sufficiently similar to the object it represents. So, for example, our illusory perception of a person must be very similar to our perception of that person in the actual world. I do not think that the illusory experience must be perceptually *indistinguishable* from perceiving the analogous object. But they must be similar enough to trick our visual system into perceiving the illusory object as if it were the real thing. This may manifest in the behavioural motivations that we have discussed. Second, the PIT implies that the object's reality is a part of our visual experience. That is, we see the object as *existing*. This follows from the CIT#3: our perceptual experiences cause us to be disposed to believe that the fictional object exists.

Many visual representations meet the first criterion for the PIT. Traditionally created

motion pictures create a physical record of actual objects. And many paintings, photographs, and drawings are remarkably life-like. But that does not necessarily mean that we perceive these objects as *existing*, as meeting the second criterion.

Carroll rejects the PIT, arguing that it our visual experiences do not meet the first criterion. Our perception of movie screens and actual objects are not identical, or even sufficiently similar to the perceptual experience of real-life objects. There are surface interferences—scratches and dirt on a film strip, hair on the projector slide, the size and shape of the screen, etc.—which make the viewer aware of the screen and remind her that the objects in the movie are not really in front of her. Carroll also points out that we typically perceive *edge phenomena*; we can see around the edge of an object as we move. We do not experience edge phenomena in our visual perception of film. We cannot look around a character to see what is going on behind her. We can give the same sort of explanation for other fictional media. We do not perceive plays in the same way that we do real-life people and events, because of the stage and other spatial and physical discrepancies between them. Pictures are always framed and are not subject to edge phenomena, just like films.[2] Even listening to an audiobook will probably not sound identical to listening to real people give an account of their lives. Each of these aspects of our experience of fiction block the first criterion of the perceptual illusion thesis.

Carroll thus rejects the perceptual illusion thesis and argues for a *recognition prompt thesis* to explain how we come to recognize objects in a motion picture. He claims that "humans acquire the capacity to recognize pictures, including moving pictures, naturally, rather than conventionally, at the same time that they acquire the capacity to recognize the objects that pictures represent"

[2] See Derrida 1987 and Foucault 1970 for interesting discussions of the relevance of frames and framing for how we experience paintings.

(*ibid*, 109). This means that we can recognize representations of real things in a motion picture; motion picture shots are "natural recognition prompts" (*ibid*, 110). We recognize the airplane in a picture of an airplane; we recognize Abe Lincoln in his portrait. We may also recognize unreal entities like vampires, not because vampires exist in our world, but because we are familiar with the concept of a vampire and, probably, have seen pictures of vampires before. Importantly, the prompt allows us to recognize the object that it represents but does not force us to perceive the fictional representation as existing. This means that our general experience of fictions may fail to meet the two criteria for the perceptual illusion thesis.

I am sympathetic to Carroll's recognition prompt thesis. Indeed, I think that something like this theory can explain how we perceive objects on a movie screen or in a picture. It can also explain the basis of our emotional and moral responses to fictional entities. I will return to this point in the following chapters. For now, though, I want to point out a potential problem for this proposal. Perhaps in most cases (maybe even all cases) there is enough similarity between a represented object and its real-life counterpart for us to recognize the representation *as* a representation. In most cases there are also sufficient visual differences between our perceptual experiences of the fictional and real-life objects, so that we do not perceptually confuse the two (barring some trompe l'oeil paintings). But there are two points worth making here. First, it is at least conceptually possible that a film could be in every way visually identical to a real-life experience. Suppose, for instance, that sometime in the near future we will be able to enter into virtual realities that are perceptually indiscernible from the actual, physical world. If this is true, then our perceptual experience meets the two criteria for a perceptual illusion.

Carroll takes pains to point out that his recognition prompt thesis concerns our everyday experiences with actual films—not a possible virtual reality like the one I just mentioned.

Nevertheless, I think that the virtual reality example highlights something missing in his critique of the illusion thesis, something that must be accounted for if we are to fully reject it. My second point is that the very recognition of fiction objects is a kind of low level sensory illusion. One could argue that, in terms of basic sensory information, our perception of visual fiction is sufficiently similar to those of real-life objects in order to motivate action. This would imply a perceptual illusion at very early stages of visual processing even if later judgments concerning the reality of the perceptual object block any kind of behavioral or cognitive reactions to the fictional object in terms of an actual thing. Neither Carroll's critique of the PIT nor the recognition prompt thesis can adequately explain this.

4. The perceptual dilemma

Imagine that you are standing along the wall of a narrow, rectangular bedroom. What do you see? The walls are a pale blue, the only window a light mint green. Several portraits line the wall next to the door. A landscape painting teeters over a large, wood-framed bed covered by a bright red blanket. Chairs, towels, and dressing tables scatter the little remaining space. Now imagine that you are at the Van Gogh Museum in Amsterdam. Standing in front of the painting *Bedroom in Arles* (1888), you encounter the same scene—only smaller, of course, and in the style of the Post-Impressionist painter. Once again, you perceive a rather awkward arrangement of furniture and wall-coverings.

What's the difference between our perception of these two scenes which, let's imagine, are

identical in terms of the objects, colors, proportions, etc.? The van Gogh painting will never trick a viewer into thinking that it really is the bedroom, due to both its size and style. Stylistic differences will act like motion pictures' edge phenomena, keeping us from seeing the painting as a real scene. But imagine that someone took a color photo of van Gogh's bedroom in Arles and made the print large enough to match the size of the actual room. What then? The main difference, of course, is on the one hand we perceive a representation of a bedroom, while on the other we see the bedroom itself. Surely we would not be tricked into thinking that we actually see the painter's bedroom. The context is all wrong, the light in the scene will never change, we can't move in that space, etc.

We do not see an actual bedroom in front of us. Yet it doesn't seem strange or out of place to say that we nevertheless *see* a chair, bed and landscape in the painting (see Lopes 2005). What do we see when we visually perceive a painting, film, cartoon? Indeed, what sorts of things do we actually perceive, in general? An investigation of this question will help us to make sense of the illusion thesis and determine whether it is plausible that we perceive visual fictional objects as real objects.

There are two questions worth addressing here. First, we want to know what kinds of objects and properties we perceive—in other words, what kind of properties we visually represent.[3] According to many theories of perception, we only perceive very basic kinds of properties, such as shapes, colors, depth, motion, location, and illumination (e.g. Clark 2000, Brogaard 2013,

[3] I assume that some version of representationalism is correct (see Byrne 2001, Harman 1990, Tye 1995, amongst others). An interesting extension of my argument here would be to examine how the illusion thesis bears on disjunctivist theories of perception (see Fish 2009, Martin 2004, McDowell 1982, etc.).

Dretske 1995,Tye 1995; see Siegel 2010 for an overview).[4] It's another question whether we actually perceive depth or illumination in a two-dimensional screen or painting as opposed to colors that merely suggest depth or illumination. So the contents of our perception of the van Gogh painting, then, only include the blueness of the walls, the greenness of the window, redness of the blanket etc. On this view, we do not perceive objects, but rather just the surface of objects (Clark 2000). Low level properties result from retinal stimulation; all other visual properties, including objects or object types, are the result of later cognitive processing—thoughts or judgments—of this basic sensory information (O'Shaughnessy 2000).

Other theorists hold that we perceive properties beyond color, shape, etc. (e.g. Bayne 2009, Peacocke 1992, Siewert 1998, Siegel 2006). On this view, we can perceive natural and/or artificial kind properties (animal, dog, or chair), causal properties (that A causes B), emotional properties (being happy or angry), and semantic properties (hearing a word's meaning). For example, Siegel argues that there are cases in which a knowledgeable subject encounters a visual stimulus of a pine tree and she sees not only an array of shapes and colors at particular locations, but also a pine tree. This may suggest that our background knowledge, beliefs, and abilities can influence our perceptual experience (Siegel 2006). As we will see, the high and low level theories of perceptual content will have different implications for the illusion thesis, but neither implies that the object of our perception exists. That is, we do not perceive things as existing or not existing.

The second question concerns whether or not there is a difference between our capacity to perceive *actual* objects and representations of objects. We literally see the paint that van Gogh used to create *Bedroom in Arles*. But do we also see a bed? There is a vast literature on the

[4] Although it is debated whether causal properties and objecthood are perceptually low or high level.

transparency of photographs which states that, due to the mechanistic nature of the production of photographs, we literally see the objects represented in a photograph (Lopes 2003 & Walton 1984; see Currie 1990 & Carroll 2008 for critiques of the transparency thesis). Photographic transparency may also carry over to the transparency of film and digital media, so that we literally see the objects being filmed/recorded. Of course, paintings and cartoons are not transparent even if photographs and film are, so some other account is required to explain how we see objects in them. Alternatively, it may just be a natural fact of our perceptual system that we see represented shapes as objects, as Carroll's recognition prompt thesis holds (see also Lopes 2005 & Wollheim 1980). A painting or film does not need to be transparent in order for us to recognize objects as being represented.

I will return to this question in the following chapter but, for now, I will remain largely silent on it. I do not think that the recognition of a represented object requires that the representation be transparent (so that we really do see through the representation to the object itself; see Walton 1984). It is enough to show that we can recognize the representation of an object on a screen or canvas. The question is whether our recognition of objects amounts to a perceptual illusion. I think that answering this question depends on how we respond to the first: which kinds of properties are represented in perception? I argue that both the high and low level positions pose serious problems for the illusion theorist. Let's begin with the idea that we perceive high level properties.

Option 1: we perceive higher-level properties

First point, what kinds of high level properties must we perceive in order for the illusion thesis to

be true? Remember that the perceptual illusion thesis claims that we are "tricked" into perceiving the visual objects of fiction as actual objects. It is not enough that we recognize fictional objects *as objects*. We also need to perceive them as actual, existing things—that they are perceptually present to us (see Dokic 2012, Noë 2006, Siegel 2010).

Here's an example. Imagine that you are walking down the street and you see a neighbor walking her collie puppy. On the current view, you perceive the puppy as belonging to a particular kind: "animal," "dog," and maybe even "collie," besides also seeing the object's shape, size, motion, color, etc. Similarly, when you are watching an old *Lassie* rerun, you may perceive the dog as an "animal," "dog," or "collie" due to visual recognitional cues, along with other low level properties. Importantly, in both the real-life and fictional cases, we perceive the dog *as a dog*. The question is whether the fact that you perceive an object at all implies that you see it *as* an actual, existing thing. This would require that we perceive the neighbor's collie as possessing a property of existing—or, similarly, that we are in its presence. The PIT further requires that we also perceive Lassie as existing, even while we know that she doesn't.

Do we have the perceptual experience of objects as a) existing or b) present to us? Let's consider the first possibility. There are ways to understand this. First, it could be that our default position is to perceive all objects as existing, as possessing the property of existence. This means that we perceive represented objects in paintings, films images, etc. as existing *by default* and we form a perceptual belief in the existence of the object. Thus, we would be under both a perceptual illusion and a cognitive illusion. Of course, other judgments, beliefs, and knowledge about the fictional status of the object may undermine this. We may discard or override the perceptual belief so that we never come to believe that the fictional entity is real. These other judgments, etc. could prevent our actually behaving as if the object exists, but not necessarily our early stages of

behavioral motivation or priming.

One issue with this interpretation is that it makes the perceptual illusion thesis trivially true. We perceive representations of objects as existing, but that's because we perceive *everything* as actually existing. We would never perceive fictional things as "being fictional" because we perceive everything as actually existing. The question is whether our perceptual abilities are actually equipped to perceive things as existing, as possessing the property of existence. Note that even philosophers of perception who do argue in favor of high level perceptual properties (such as Susanna Siegel) do not claim that we perceive complex properties like "existing." Indeed, the most these philosophers claim is that we can perceive *kind* properties. If so, then we do not, strictly speaking, perceive objects as existing or not existing; we *judge* them to be so.

The second possibility complements this last claim. We do not perceive an object possessing properties like "existing" or "nonexistent." Attributing those properties to an object must be the result of cognitive processing, not perceptual processing. On both interpretations, the perceptual system does not distinguish between fictional and non-fictional objects. The difference is that, on the former view, we default by perceiving the object as actually existing. On the latter, we default by suspending judgment as to the ontological status of the object, because existence is not the kind of property that we can perceive without cognitive mediation. In both cases, we may be motivated to act based on our perception due to direct behavior links that do not require cognitive processing, such as a cognitive belief.

The difference between these two positions is subtle, but quite important. On the second position, cognitive processing is required to interpret the object as non-existent, as a representation on a canvas or screen. Perception alone cannot indicate that something is non-actual.

The "no distinction in perception" position implies that the PIT fails, at least on the reading

that we perceive that fictional objects exist. It's possible, however, that we perceive fictional objects as "in our presence"—as in front of us spatially, just as we would real things. This seems unlikely. On most accounts of perceptual presence, we are perceptually present with an object if that object is in our egocentric space (Dokic 2012, Noë 2006 and Siegel 2010; see also Evans 1982 on direct reference). This means that we are disposed to act in certain ways towards the object, and such actions will affect how we perceive it. For example, a tomato is perceptually present to us if we can orient our body in relation to it. This isn't the case for visually represented objects; as Carroll notes, represented objects lack edge phenomena and other egocentric perceptual capacities.

I conclude that we do not perceive fictional objects as existing; our visual experience of fictions does not meet the second criterion of the perceptual illusion thesis. And if we do not *perceive* fictional objects as existing, then we will not be disposed to cognitively believe that they exist either. That means that CIT#3 fails as well.

Option 2: we only perceive low level properties

Perhaps the question of whether we can perceive objects as existing is a moot point. After all, many philosophers of perception argue that we only perceive low level properties (color, shape, location, illumination, motion, depth, etc.), not high level ones and certainly not properties like "existing." This means that thoughts and judgments that allow us to process perceptual information as objects of particular kinds (Stokes 2014).

Let's illustrate this in terms of the Lassie example. When I'm walking down the street and I see my neighbor's collie, all that I actually perceive is a particular color, shape, size, motion, and location in my visual field. Later cognitive processing puts this information together and informs my experience of it; now I see that this thing is a dog that is walking towards me and, if I'm a dog-

sophisticate, that this particular dog is a collie. But properties like "dog," "collie," and even "object" may not feature into my basic perceptual experience. Now consider my perception of Lassie on TV. Here, too, I perceive an image in front of me that transmits information like color, shape, motion, and location. I do not represent the properties of dog, collie, or Lassie until more cognitive processing has taken place.

So where is the perceptual illusion? On this position, there is none! We do not perceive objects, real or unreal. We only judge, in some way, that we see an object. In other words, our perception in this cases passes the first criterion for the perceptual illusion, but fails to meet the second. It is very likely that our perception will be informed by other cognitive beliefs and concepts that indicate that this thing we see on the television screen is a fictional entity. An exception to this would be if we are completely ignorant of whether the image we see is real or not, as in a perfectly indiscriminable virtual reality. I am willing to grant that this would be a perceptual illusion.

5. Cognitive penetrability & the illusion thesis

It is possible that the perception of high level properties is mediated by top-down beliefs and judgments—specifically, knowledge about the fictional status of the object we perceive. If so, then it may also be possible that our beliefs about the nature and properties of an object may shape how we perceive it. On this view, we might actually perceive objects as existing or not existing because of a cognitive mediation from beliefs and judgments about the ontological status of the object. One basis for the cognitive judgment that a fictional object is nonexistent stems from the visual discrepancies between visual representations and real-life objects that we discussed above: two-

dimensionality, screen/canvas/photographic obstructions, etc. These differences cause us to judge that the object we perceive does not actually exist. That judgment, in turn, shapes how we perceive what is represented on the screen, in the sense that our experience is not sufficiently similar to perceiving a real-life object (criterion 1) and that we do not see the object as existing (criterion 2).

This would mean that our perception of a visual representation is *cognitively penetrable*; our thoughts, beliefs, and knowledge about an object may influence how we actually perceive it. This is an interesting, though controversial, possibility. If true, it may be possible that we perceive fictional things differently than we perceive real things. Combined with Carroll's points concerning the visual differences between visual fictions and real objects, the possibility of perceiving fictional entities as real objects never arises in our everyday experiences.

The cognitive penetrability thesis holds that our knowledge, beliefs, and thoughts about an object can influence how we perceive it. I address this issue separately since cognitive penetrability can apply to either high or low level perceptual properties. Importantly, if cognitive penetration of perceptual experience actually occurs, then our beliefs and thoughts about the status of the fictional object might help us to literally see the object as not actual. I will not argue for or against the cognitive penetrability of perception. My goal here is to simply explore its implications for our perception of visual representations.

Carroll's account of the perceptual illusion theory implies that perception is cognitively impenetrable.[5] This means that, no matter what we believe or know about an object or illusion, our perception of it will remain the same. The dog expert and novice will have the same visual experience of the collie. In contrast, Siegel states that:

[5] Carroll borrows this notion from Zenon Pylyshyn's arguments on cognitive impenetrability (Pylyshyn 1999). See also Jerry Fodor's modularity thesis (1983).

> if visual experience is cognitively penetrable, then it is nomonologically possible for two subjects (or for one subject in difference counterfactual circumstances, or at a different time) to have visual experiences with different contents while seeing and attending to the same distal stimuli under the same external conditions, as a result of differences in other (including affective) states (Siegel 2012, 4).

Dustin Stokes (2014) argues that cognitive penetrability has interesting implications for our evaluation of artworks. Someone who possesses a great deal of knowledge about an artwork (its artist, historical context, genre, etc.) will perceive the work differently than someone who lacks that knowledge. Information-rich perceptions of an artwork help us to evaluate it; the informed viewer may be capable of making more appropriate aesthetic evaluations than the novice.

I am neutral on Stokes evaluative claim. I am more interested in the case he makes for how cognitive penetrability impacts our perception of both and high and low level properties. Perceptible high level aesthetic properties include the "standard aesthetic properties of being graceful, serene, vivid, or delicate (*ibid*, 15). The expert's internalized knowledge of a work's category, historical context, and artist, will affect her overall perceptual experience so that she perceives it in terms of these aesthetic properties, as well as others. She will perceive Picasso's *Guernica* not just as a painting with certain colors and shapes, but also as violent, disturbing, or sorrowful. Stokes suggests that it is also possible for the expert to perceive additional high level properties that pertain to the work's category: "One may just see the impressionism or just hear the Brahmsianism in a work" (*ibid* 16). In both cases, the expert's knowledge of the work influences her perception of it.

Stokes makes a similar argument concerning the cognitive penetration of low level perceptual properties (for those skeptical of high level property perception). What we know about an object may change how we perceive its color, shape, or size. For example, we associate certain

objects with stereotypical colors (a blue Smurf, a red Coca-Cola icon, a yellow banana).[6] We expect to see these objects as being a certain color, so when we are put in a situation in which we have to experimentally match these objects with a color, we tend to identify it with the *expected* color. This occurs even when this particular object actually possesses a unique color-property, like a pink Smurf instead of the standard blue (Siegel 2011). Stokes (and others) takes this as evidence for cognitive penetrability effecting our perception of low level properties; our belief that a Smurf is typically blue shapes how we see Smurfs. We will expect them to be blue and perhaps even perceive them as blue when they are not.

The same basic principle can apply to artworks. It could be that an expert perceives the organization of low level properties differently than the novice—it's not that they see a different color, but rather that the expert sees how those properties are organized while the novice does not (Stokes 2014, 21). For example, the expert will see the relation between the lines and colored rectangles of a Mondrian *Composition* piece differently than a novice will. The expert sees the same colors and shapes as the novice, but also the "lack of negative space and the dominance of colored rectangles," because the expert is familiar with Mondrian's oeuvre and the novice isn't (*ibid*, 21). Stokes also suggests that it is possible that the actual low level properties themselves might be differently perceived by the expert and the novice. We learn to associate certain colors with certain paintings (bright red for Matisse, blues and grays for Picasso's blue period, pinks and creams for his rose period, black and white for Motherwell, etc.). These learned associations may influence how we expect to see a work, just like the blue Smurf. In Stokes' example, a Rothko expert who is familiar with the painter's monochromatic multiforms, such as the all-black paintings in The Rothko Chapel, will likely perceive subtle variations in each painting—unusual

[6] See Witzel et al 2011 for the empirical study. See Deroy 2013 for a discussion of these studies.

patterns in the all black canvas, for example—whereas the novice will not, even though they are both looking at the exact same painting that has the exact same properties.[7]

The cognitive penetrability of perceptual experience may motivate the idea that knowledge that the object of our perception is a *nonexistent* object, rather than an actual one, may influence how we perceive it—both in terms of what low level properties we notice, expect, and perceive as well as how we categorize the image. For example, an expert may notice aspects of van Gogh's *Bedroom in Arles* that make her perceive the image as being impressionistic, highlighting the visual discrepancy between the representation of the objects in the painting and how they would appear in real life. Someone who is competent at watching film fictions may note that the use of sound and shade in a film-noir are overly dramatic, much more so than they would be in real life.

Does this mean that we perceive the property of "being fictional"? I do not think so. It does mean, however, that our perception of fictional entities will be slightly different than our perception of actual entities, because of the properties that we are inclined to notice. Experts may have the capacity to perceive both high level properties of a fictional object and low level properties in a different way due to their knowledge of the fictional status of the object. This suggests that the illusion would not obtain in such cases, if the cognitive penetration of perception is possible. The expert may not perceive fictional entities as actual ones. The differences between perceiving an object as fictional or actual may be subtle (and maybe unconscious), but they will nevertheless be there. Our background knowledge that the object we perceive is fictional may shape how we perceive it. If that's true, then we may not fall prey to a perceptual illusion. Our

[7] It's important to distinguish a genuine case of cognitive penetration from that of perceptual learning, according to which one can learn to recognize particular properties and objects. In a case of cognitive penetration, the subject directly perceives a property in a different way due to her knowledge of a particular object—so, for instance, she sees the Matisse painting as redder than she would if she did not know about Matisse's work and associate it with the color red.

knowledge that an object is fictional will make us perceive that object differently than we would if we were actually in its presence.

6. Resisting the illusion thesis

Let's return to the three versions of the cognitive illusion thesis that we encountered in §2. They state:

> CIT#1). We have a conscious, occurrent belief that the objects of fiction actually exist.
> CIT#2). We have an *unconscious* belief that the objects of fiction actually exist.
> CIT#3). We have a perceptual belief in the existence of fictional objects.

I have argued that the standard functionalist/phenomenological responses to the CIT provide sufficient evidence to block CIT #1 and CIT#2. We generally lack both conscious and unconscious beliefs in the existence of fictional objects. If we did, we would likely act in different ways towards fictions than we actually do.

However, the behavioral evidence left open the possibility that we have an perceptual belief in the existence of fictional objects. This could mean that we are, in fact, motivated to act based on our belief in the existence or presence of a fictional entity. CIT#3 rests on the further claim that we perceive fictional entities as existing—in other words, that we fall prey to a perceptual illusion concerning the actuality of a fictional entity. I have argued that the PIT is largely a question concerning the types of content that we can perceive: high level or low level properties. But neither view suggests that we can perceive objects as existing or not. Finally, research concerning the

cognitive penetrability of perception, if possible, may also support my rejection of the perceptual illusion thesis. It is possible that our thoughts and beliefs concerning the existence of fictional entities actually effect how we perceive them, as fictional.

I will now briefly sketch my alternative explanation of the low level behavioral and emotional data, based on what I called the "no distinction in perception" view in §4. On my view, perception itself is silent concerning the actuality of an object; it only allows us to recognize objects. This means that there really is no perceptual illusion thesis; if there is an illusion, it comes in at the cognitive level of belief. It is more likely that our perception of fictional entities is neutral concerning the objects' existence and that later judgments (or top-down beliefs and thoughts) later come into play to determine their existence. Generally speaking, these judgments will determine that the fictional entity is, in fact, fictional.

This view further implies that unconscious, automatic dispositions to act are not actually based on beliefs, but rather stem straightforwardly from perceptual processes. We may perceive the fictional entity as being a certain type of object, but we do not form a perceptual belief that the object is real, that it actually *exists*. Basic perceptual processes (combined, perhaps, with some associative processing or inferences from stored concepts, as well as cognitive judgments) allow us to recognize objects in visual fictions. Perceiving something as a being particular object may prime certain actions—that is, there may be a direct link between the perceptual processes involved in object recognition to action priming. For example, I will behaviorally respond to a coiled up piece of rope if I recognize that object as being snake-like. This does not entail that I perceptually believe that I am in the presence of a snake. The perceptual system does not distinguish between existing and non-existing objects on this view, but rather will respond to both similarly; I will defensively react to both the coiled up rope and a snake. Later, cognitively meditated judgments

may prevent some primed actions from arising. I may step around the coiled object rather than flee once I realize that there is no snake in front of me.

I think that this position can make sense of our behavioral and psychological responses towards fiction. Perhaps the reason why we fear fictional monsters, snarling dogs, and serial killers is because we perceive them in much the same way that we would the real thing—not as perceptually identical to the object, but as possessing enough recognitional cues for us to visually recognize it. Some quick, automatic behavioral responses to these objects are the direct result of our perception, before later judgments kick in to influence them. This could mean that we have automatic affective responses to a fictional entity before we have another belief (conscious or unconscious) to the effect that the entity is not real. We would be primed to respond to fictional entities, at least until later cognitive processing kicks in to halt those behaviors. On the other hand, slower, more complex actions—like getting up from one's seat and leaving the theater—would be blocked by the judgment that an object is fictional.

As a result, we also do not form a perceptual belief that what we perceive in a fiction is real, because we do not perceive it as such. But what about our low level behavioral priming? One way to interpret this data suggests that we are primed for action when we perceive a character's movement in a painting or on a screen even though we consciously know that the character is not real. It could be that these areas of the brain—the so-called mirror neurons in motor cortex—are not influenced by top-down knowledge, unlike visual perceptual areas. Thus we will often be primed to act even when we do not believe that the objects we perceive are real things with which we can interact. Again, this means that action priming may arise directly from our perceptual experience. It's likely that we automatically react to certain value-laden objects in our environment even before we make any judgment concerning that object's existence. This would be a good

survival strategy. Behavioral priming results from purely brute, causal recognitional cues; anything remotely resembling a person or animal (for example) may prime us for action in one way or another.

I resist calling this a cognitive illusion; belief does not seem to play a role in the perceptual/recognitional capacities that prime us for action. This seems to be the case for our affective reactions which may lead to behavior even without this neural representation reaching cortical areas of the brain that mediate beliefs (see LeDoux 1996 & 2012; Damasio 1994 for just some evidence of this). This basic framework may also apply for non-emotional perceptual processes as well. Of course, one might say that these low level responses are irrational even if they are not illusory, because they contradict what we consciously know and endorse about the world. Better yet, they are *a*rational because they are not shaped by cognition. We will be primed to respond to certain objects whether we believe that these things are real or not. Generally we do not believe that fictional entities are real. As I suggested above, our beliefs coincide with our actual behaviors, or, as the case may be, how we do not behave. If we do not act as if a fictional character is real, then we probably do not believe that it is.

I have spent some time reviewing and arguing against the illusion theses in order to thwart a potential worry for the SAV. One could argue that illusory perceptions or beliefs are sufficiently different from non-illusory ones to constitute a different type of mental state. I have attempted to show that this worry is unfounded. Even if the illusion thesis is true, it does not entail that the distinct attitude view is correct. We perceive fictional objects in the same way—by the same mechanisms and in terms of the same broad types of mental states—as we do actual things. We do not need to posit a distinct cognitive attitude to explain our perceptual experiences of fictions.

Chapter 3: Taking the Fictional Stance

1. Two foundational questions of fiction

Imagine watching a local performance of *Much Ado About Nothing.* You see actors playing Benedict, Beatrice, Claudio, and Hero, the scaffolding representing an Italian villa, and papier-mâché rocks and trees portraying the country landscape. Whether in a bustling theater or lounging on a park lawn, what we see on a stage is nothing more than this: scaffolding, papier-mâché, fake wooden tables and chairs, and actors. But it also seems like we "see" something more than the mere stuff on a stage: the fictional world of the play.

There is a sense in which the second type of seeing is perceptual; I *see* Benedict and Beatrice. How is this possible, if physically, there are only actors on the stage, and not fictional entities? I contend that "fictional seeing" of this sort involves a kind of transformation. Of course, we cannot physically transform a papier-mâché tree into a fictional one. The transformation must occur mentally. There is much debate concerning the type of mental activity is involved in an audience's capacity to see physical materials as the fictional entities they represent. It requires the capacity for object recognition: we recognize the objects on a stage, screen, or painting as people, tables, trees, etc. But the capacity to see things as objects is not enough for us to see the actors on a stage *as* the fictional entities, Benedict and Beatrice. For what we seem to see does not actually, currently exist. How is this transformation possible? Call this *the question of fictional transformation*.

Another issue arises here. We generally know that the object of our engagement isn't real. How, then, do we take the objects of fiction as the types of things that we can respond to and

judge—emotionally, morally, and otherwise? In other words, why is it that we have mental attitudes towards the objects we encounter in fiction? Call this *the question of fictional response*.

Together, these two questions capture the foundational characteristics of our psychological interactions with fiction: how we understand objects as fictional objects, and why we respond to them as we do. Ideally, we can develop answers to both the question of fictional transformation and the question of fictional response with one theoretical framework, as other philosophers have attempted (see, for instance, Currie 1990 and Walton 1990). My goal in this chapter is to do just that.

Let's begin by considering several possible answers to the two questions. We encountered these positions in chapter 1; I will discuss them in more detail here. First, it may be that we *pretend* that the objects we perceive onstage are fictional characters, places, and things (see Searle 1975 & Kripke 2013). Our mental states about fictional entities are pretend mental states; we *pretend* to believe that Benedict secretly loves Beatrice, but we do not *actually* believe this. Likewise, we pretend that we see Benedict and Beatrice, but we know that we don't actually do so. Second, we may just *imagine* that we see fictional objects before us; we imagine that something is the case when, in fact, it is not (see Weinberg & Meskin 2006; see Van Leeuwen 2013 for more on different types of imagining). Finally, *theories of make-believe* involve the application of both pretense and imagining. According to these theories, we make-believe that the objects on stage are actual in a fictional world (see Currie 1990 & Walton 1990). Theories of both imagination and make-believe posit that we have imaginary mental states about fictional stories (e.g. we have an imaginary belief that Benedict loves Beatrice).

Pretense theories, imaginative theories, and theories of make-believe each posit that we utilize mental states during our engagements with fiction that are different in kind from those used

during our everyday interactions with people and things. These theorists also generally hold that pretense, imagination, etc. are a part of other capacities as well, such as pretend play and hypothetical thought. I will focus here on our interactions with fictions, although many of my arguments will transcend to these other areas. The distinct attitudes allow us to mentally transform the actual objects we encounter into fictional ones. They also explain how we can mentally respond to fictional entities that we know are not real. As we have seen, I call this general position *the distinct attitude view* (DAV). Our emotions, beliefs, judgments, and desires about fictional objects can be easily explained insofar as we pretend (etc.) that these are the types of objects that we can respond to. We have *pretend* mental responses to *fictional* entities.

I argued against the main motivations behind the DAV in chapter 1. One concern with the DAV is that many accounts build their theory on a flawed notion of the nature of the mental states in question. Once we have a clear grasp on the functional role and ontology of mental states then it is unlikely that we will find a distinct attitude warranted to begin with. Furthermore, one can argue from a principle of parsimony that there is no need to posit distinct mental attitudes if ordinary ones have the same explanatory power. I will attempt to show that we simply do not need to posit a DAV to answer the two foundational questions of fiction. Finally, the DAV cannot account for our phenomenological—that is, conscious, possibly introspectable—experiences with fictions. Our emotions, beliefs, desires, and other mental states towards fictions feel natural and relatively automatic, not like we are playing a game of make-believe, simulating, or even imagining a possible course of action. In fact, our mental states about fictions often do not seem any different than those about actual things.

Because of worries like these, I argue that the answers to the questions of fictional transformation and response are compatible with a *standard attitude view* (SAV). This is the idea

that the mental state types involved in our interactions with fiction are not unique to those contexts, but rather are of the same type as those in ordinary, real-life contexts. We have standard beliefs, emotions, desires, etc. *about* fictional entities and states of affairs. Rather than positing unique mental states for our interactions with fictions, I will argue that we can explain the two questions in terms of a difference in the intentional objects of our mental states (i.e. fictional objects vs. real-life ones).

The SAV can take many forms. In what follows, I will develop my own version: *the fictional stance*. In brief, we take the fictional stance when we recognize that the object of our engagement is fictional. Doing so allows us to see representations of fictional entities *as* fictional entities. We do not pretend or imagine that Benedict is on the stage before us; we (in some sense) *see* and *think of* him there. Taking the fictional stance also means that we treat fictional objects as the kinds of things that are appropriate objects of our mental engagement. By taking the fictional stance, we see the fictional objects as the real objects that they represent, objects that we would normally pity, care for, feel with, and judge. The fictional stance does not count as an instance of the DAV. Rather than explaining our behavioral and attitudinal responses towards fictions in terms of distinct mental state types, I explain them in terms of a particular kind of intentional content. In other words, we utilize the same types of mental states and mechanisms during our engagements with fictions that we do in our everyday lives, but these mental states are about fictional objects.

The following sections flesh out the fictional stance as an alternative to the DAV. The arguments in each section are cumulative, building off each other until we garner a complete picture of our psychological engagements with fictions. The full explanation of the fictional stance will not be complete until §4. The central tenets are as follows:

1.) We know that the representation meets three conditions for being a fiction: is non-

actual, is created, and depends on particular objects and people in order to persist.

2.) We can recognize representations as of or about particular kinds of objects.

3.) We recognize these represented objects as being fictional objects.

I will present each of these points in §3, thus establishing the foundation of the fictional stance, after critiquing the DAV positions in §2. My own view builds off a commonsense ontology of fiction, an account of representational seeing, and a variation on Arthur Danto's concept of the 'is' of artistic identification (Danto 1964). None of these individual positions can answer the two foundational questions on its own. Nevertheless, together they form the foundation for the fictional stance which, I argue, can both answer the two questions and serve as a general framework for a psychology of fiction. In §4, I will lay out the fictional stance and show how it answers the questions of fictional transformation and response. I will consider two other versions of the fictional stance in §5. §6 applies the fictional stance to several ubiquitous metaphysical and semantic questions concerning fictions. I conclude with a brief comment on the explanatory power of the fictional stance.

2. Pretense, make-believe, & imagining

I have argued that we are under neither a perceptual nor cognitive illusion while engaged with visual fictions. A more plausible possibility is that we *pretend, make believe,* or *imagine* that fictional objects are real. These terms are often used interchangeably in the literature. I suggested in the first chapter that doing so is a mistake. As we will see, we can differentiate between these types of mental activity in various ways. I will attempt to make these distinctions clearer in this section.

The DAV can be cashed out in terms of one of two claims. First, we want to know whether there is a unique mental state that is utilized solely in fictional contexts—a pretend or imaginative belief, for example, that is similar to, but cannot be identified with, a stereotypical belief. Second, we want to know whether pretense, imagining, or make-believe are necessary to make of what a narrative *fictional*. I deny both claims. But it is worth noting that one can deny the first, concerning mental state types, without denying the second. The idea there would be that imagining, for example, is not a *mental state type* that is unique to fiction, but it still is a necessary part of a engaging with fictions. This would make our engagements with fictions distinct from our engagements with non-fictions. We need to deny both claims in order to properly combat the DAV.

I will consider the different versions of the DAV in turn. We encountered the pretense thesis back in the first chapter. This view states that an appreciator of fiction pretends that the fictional story, characters, and world are real for the duration of her engagement with it. This position has important ramifications for the ontology and semantics of fiction. Fictional objects, places, and characters are not actual things (either concrete objects or abstracta). Rather, we merely pretend that they are real, like the objects of pretend play. We speak, feel, and think about fictional entities *as if* they are real things, but we know that they are not. This view makes the most sense in terms of dramatic art. Actors pretend to embody the character that they portray. In so doing, they represent a fictional story; for instance, the development of a romance between Benedict and Beatrice in *Much Ado About Nothing*. The actors pretend that they are certain fictional characters and the audience joins in their pretense. We pretend that we see real events and people, even though they are actually fictional *representations* of events and people.

Semantic considerations may be the greatest motivation behind pretense theories. Pretense theories state that speech acts about fictional entities are neither genuine assertions of propositions

nor do they refer to actual objects. Rather, we pretend to make assertions about fictions and pretend to refer to nonexistent things. Saul Kripke (2013) claims that pretense is also involved in the composition of fictions. An author pretends to refer to an actual person when she creates a character. The character itself can either be understood as an abstract entity, possible object, or definite description. In each case, when we say the name 'Benedict' or 'Beatrice' we pretend to refer to a real, existing person. For some sentences, we may refer to the abstract entity, possible object, or utilize a definite description. This is the case when I say "Emma Woodhouse was created by Jane Austen." But for other sentences, we merely *pretend* to refer to a character. The author pretends that the name refers and the reader/audience goes along with this pretense. Kripke calls this *the pretense principle*, arguing that it applies to sentences internal to a fictional story, such as "Beatrice is in love with Benedict, but she won't admit it to herself." According to Kripke, the pretense principle holds no matter what one thinks about the nature of fictional propositions or fictional names (*ibid*, 24).

Pretense theory can also explain the two questions we raised in the first section: the question of fictional transformation and the question of fictional response. Our response to the first question will differ slightly depending on the art form. When we engage with a stage production of *Much Ado*, we pretend that the actual, physical objects on stage are the real thing—the papier-mâché boulders are real boulders and the living, breathing actors are the characters in the story. When we watch a film production of the play, such as Kenneth Branaugh's 1993 version, we pretend that the images we see on the screen are actually the Italian villa and countryside as opposed to wherever the film was shot. We pretend that Branaugh is Benedict and Emma Thompson is Beatrice, even though we know that they aren't really. Comparable stories could be told for video games and opera, and maybe even fictional representations that we encounter in

pictures and photographs (though I will not focus on those art forms here).

Pretense theorists can answer the question of fictional response in a similar way. Because we pretend that fictional entities are real ones, we emotionally respond to, morally judge, believe, and desire things about them in much the same way we would if they were real. These responses will always be in the context of the pretense and so will not engender the same kinds of behavioral responses as they typically would. Our mental states are run "offline," disconnected from their standard functional role.

So pretense theories seem to have a great deal of explanatory power when it comes to our psychological interactions with and speech acts about fictions. Still, many philosophers have argued against pure-pretense theories and have instead incorporated pretense into more sophisticated theoretical frameworks. I will not review these argument here (but see Currie 1990, Walton 1990). Instead, I will review general arguments against pretense theories, arguments that apply to other versions of the DAV as well.

One approach that utilizes pretense is a theory of *make-believe*. Theories of make-believe hold that our engagement with fiction is a part of an elaborate game of make-believe, similar to games we play as young children (Currie 1995, Walton 1990). These games involve an element of pretense, but also another, distinct kind of mental activity: imagination.

I will focus on Kendall Walton's theory of make-believe from his text *Mimesis as Make-believe* (1990). This remains the most detailed, comprehensive, and influential version of the view. Walton argues that *all* representations involve make-believe, including fictions. An important application of make-believe concerns a definition of what it means for a representation to be *fictional*. In typical games of make-believe, the subject stipulates certain rules concerning objects that we take to be a part of the game. For example, when children pretend-play house, they stipulate

that a mud pie is a cherry pie; the mud pie in a box is a cherry pie in an oven; the tree house in the backyard is a cottage in the woods.

Derek Matravers (2014) explains how this applies to fiction. One makes a stipulation that there is some truth about the actual world that, somehow, will take us to a truth about a fictional world (eg. there is a glob of mud → there is a pie; there is an actor → Henry V). But it is not enough that we create or understand this stipulation. Walton argues that we also need a particular kind of mental activity to get us from the actual to fictional world: imagination. Stipulated truths about the actual world are *prescriptions to imagine*—or, alternatively, *principles of generation*, propositions that mandate the reader/audience to engage in a game of make-believe. Audiences understand fictional representations, including both verbal and depictive representations to serve as props in games of make-believe. They are objects that call for us to imagine the fictional story, respond to it, and even think of ourselves as a part of it.

Matravers describes Walton's position in terms of two criteria. Each describe the imagination's role in transforming an actual object or proposition into a fictional one. First, the *transformation criterion* states that "something is a fiction if the imagination is required to transform a proposition true in the actual world into a different proposition true in the fictional world" (Matravers 2014, 13). During the game of make-believe, we imagine that the objects in a dramatic representation, the images in a picture or on a screen, or the propositions in a novel each generate fictional counterparts. We make-believe that the stage actor Ian McKlellan is Macbeth on the stage, that Kenneth Branaugh is Benedict in the film, or that the propositions in *Much Ado* make genuine assertions about a real person named Beatrice.

Second, Walton argues for the *engagement criterion*: "something is fictional if it engages our imagination as only the imagination can account for facts concerning our engagements with

fiction" (Matravers 2014, 16). The idea is that an imaginative game of make-believe is required to explain the perspective that readers/audiences take of the fiction as well as the vivacity of our psychological responses to it. Walton argues that we do not merely imagine the objects in a fiction. We also imagine *ourselves* in the fiction, as a part of the fictional world that observes the action (Walton 1990, 29). This helps to explain why our affective responses, judgments, and simply our engagement itself feels so intense—or, as Matravers puts it, why it possesses such vivacity.

These two criteria answer the question of fictional transformation and even suggest an answer to the question of fictional response. We use our imagination during the game of make-believe in order to transform the objects on a stage or on a screen into the actual objects in a fictional story. Just as children imaginatively transform a mud pie into a cherry pie during their games, so do adults imaginatively transform McKlellan into Macbeth and Branaugh is into Benedict.

Our intense participation in games of make-believe leads to equally intense psychological responses to fictional objects: emotional responses to fictional characters, judgments about them, desires concerning their well-being, etc. According to Walton, though, the mental states that function within the game of make-believe are not genuine mental states, because they lack behavioral motivation. This is so even though they *feel* the same as non-imaginative mental states (Walton 1990). The audiences' speech acts (e.g. "Watch out for the slime!" or "Mr. Knightley is so charming"), emotional responses (fear towards a fictional villain or compassion for a protagonist), and other mental attitudes (*believing* that *Firefly*'s Captain Reynolds is a good person despite his gruff exterior; *wanting* Desdemona to survive even if we know that she won't, etc.) should all be understood as acts of pretense within the world of make-believe.

We should compare our psychological engagements with fiction to a child's pretend-play.

A child creates a world of make-believe around props in her environment and she acts and speaks from within that world. A couch and set of chairs become a dungeon, a box becomes a house, a high heeled shoe becomes a ruby slipper (see also Harris 2000). Indeed, Walton suggests that one of the reasons why fictions are so important and entertaining for adults is because they are an extension of our childhood pretending. Creating such a rich make-believe world requires a fair amount of imaginative labor from both the audience and creator, but it's an ability that is readily undertaken and mastered by both adults and children.

Theories of make-believe have a certain appeal. They can explain our mental states towards fictional entities, our speech acts about them, and even the nature of those entities. Nevertheless, some of the implications of these theories have been called into question. For example, some philosophers argue, like me, that pretend or imaginary states—i.e. non-standard ones—should be resisted (see especially Matravers 2014 and Carroll 2008). Furthermore, several philosophers have argued that theories of make-believe cannot capture our actual first-person experiences with fictional stories; it does not feel as if we are involved in a game of make-believe (Carroll 1990; Neill 1993).

There is one more view to consider: that we imagine that fictions and fictional entities are real, but without pretending or make-believing that they are. On the imagination view, watching a stage production of *Much Ado* involves imagining that there is a world in which Benedict falls in love with Beatrice. This is a different kind of mental activity from pretense. Pretense involves taking some state of affairs to be true or some object to exist in the actual world. As I noted in the first chapter, imagining does not necessarily have this tie to truth. Imagining may be more like entertaining a proposition, or supposing that something is true, without thinking that it actually is or merely considering something not present. We can imagine what it would be like for some state

of affairs to occur without thinking that it *actually* does, either in this world or a fictional one. This applies to *attitudinal* (or, *propositional*) imagining in which imagining that something is the case is analogous to believing that it is the case.

Some imagination theorists contend that our imaginings about fiction constitute a distinct attitude. I imagine that there is a world in which an Italian lord visits an old friend and marries his daughter when I watch *Much Ado*. My mental attitudes are imaginative. I seem to have genuine beliefs about Henry, desire things for him, and emotionally respond to him. But imaginative beliefs (etc.) are not genuine beliefs, again, because of the distinction in motivation for behavior.

We must be careful here. I am not denying that we *ever* use our imagination when engaged with fictions. Nor am I denying that imagining is a different kind of mental act than believing. Clearly they are different. What I do deny, however, is that there are distinct types of states called "imaginative beliefs," "imaginative emotions," etc. I also deny that propositional imagining, make-believing, etc. is *necessary* for our engagement with fiction, even though we may sometimes take up these mental activities.[8] Unfortunately, not all philosophers are careful to make these distinctions.

Let's consider several general arguments against pretense, make-believe, and imagining. First, it is questionable that these theories can accurately explain our conscious experience of fictions. Alex Neill (1991) has argued that anyone who plays a game of make-believe *knows* that she is doing is. Children make conscious stipulations concerning actual objects and are aware of the resulting make-believe world in which they insert themselves. The same may go for our engagements with fictions; if we begin a game of make-believe while engaged with a fiction, then

[8] It may be that imagistic imagining is necessary for our engagements with fictions. But this hardly counts as a distinct type of state; we imagistically imagine all the time, in all sorts of contexts. See chapter 1 for more on the different types of imagining.

we will know that we are doing so. The problem is that most people would probably deny that they are, in fact, "make-believing" or "playing pretend" or even "imagining that something is the case" while watching a film, attending a play, or reading a novel. Their conscious experience of the fictional story will generally be more loosely experienced than this, lacking concrete stipulations or rules to generate imagining. In fact, it often *feels like* we are in the direct presence of fictional entities, or at least their representation—not that we have created a fictional world of make-believe (see Wilson 2011).

A reader or audience member may, if pressed, concede that she imagines or pretends that that something fictional is actually true. But we generally do not have this conscious experience during the time of our engagement with the story. This is especially persuasive if we consider our psychological response towards fictions, which *feel* the same as any typical emotion, desire, belief, etc. This kind of phenomenological point is not a knock-down argument against the DAV, for it could very well be that games of make-believe, pretense, or imaginings typically occur without conscious initiation or continuation. Still, one would suspect that if make-believe, pretense, or imagining is a necessary part of our engagement with fictions then it would at least *sometimes* be consciously experienced by *some* people. And that often doesn't seem to be the case for a general experience of fictions.

A second critique of the DAV concerns our reference to fictional entities and the use of fictional names. Recall Kripke's pretense principle: we make pretend assertions about fictions and pretend to refer to fictional entities, when in fact there is nothing to assert and nothing to which we can refer. There is a question, though, about whether the pretense principle can really do the work that the DAV needs it to. Consider a point raised by Nicholas Wolterstorff (1980): pretense theory puts the "fictioneer" (the author, playwright, filmmakers, etc.) in a role of pretending to

make an assertion and pretending to refer. Who, for example, does Nikolai Gogol refer to when he writes "Chichikov set out to look the town over"? One might naturally think that it is Chichikov, who or whatever that might be. But why think that pretense escapes the referential problem? Why can we *pretend* to refer to something that doesn't exist if we can't *actually* do so? What's so special about pretense, such that it bridges this ontological gap? Wolterstorff's point seems to be that any kind of reference—whether actual or pretend—requires that its object exists *in some way*.

I leave this as an open challenge for the pretense/imaginative views concerning fictional names. It need not be the case that these theories utilize the pretense principle, although Walton's own view (amongst others) maintains that we make pretend assertions and references to fictional entities.

Finally, we can ask whether pretense, make-believe, or imaging *entails* a distinct attitude. On one reading, they do. By definition, pretending, making-believe, and imagining are distinct kinds of activity that we use during our engagements with fictions, but not in real-life. We pretend (etc.) that an image is a person; we do not (generally) pretend anything about the nature of a real person. On another reading, however, it could be that these views to not require a distinct attitude. It could be that we pretend that fictions are real while still maintaining that throughout this pretense we have genuine mental states that are processed in roughly the same way that a genuine state would be (barring differences in content and keeping in mind my points concerning motivation to act). The idea would be that I pretend that Hamlet is real while genuinely believing that he is the prince of Denmark, that his mother was involved in his father's death, etc.

One might still worry that imagining, pretending, etc. mark a distinctive activity that is used solely in our engagements with fiction. But this is not the case. We imagine, pretend, and make-believe about actual or possible objects, as well as fictional ones. Derek Matravers makes a

relevant point here (Matravers 2014). It is not only fictions that stimulate pretense, imagining, and make-believe. These activities can be stimulated by historical narratives, news broadcasts, and biographies in the same way they would by fictional stories. Furthermore, we attitudinally imagine outcomes for day to day decisions: what we will have for dinner this evening, where we would like to attend graduate school, what kind of car we would like to purchase. We might even pretend that we have a new fancy car or make-believe that our current clunker is a Porsche. These mental activities are not strictly relegated to fiction and so they are insufficient as general accounts of our psychological experience of them. Matravers argues that if there *is* a distinct attitude at work in fiction, then that attitude *also* is at work in our engagements with narrative representations *in general* as opposed to just fictions.

My own view contends that not only are imagining, pretense, or make-believe *insufficient* for our experiences with fictions, they also are *unnecessary* for those experiences. We can plausibly explain our engagements with fictions without appealing to these notions. It's not that these capacities are never involved in our engagements with fiction. Rather, I argue that they do not *need* to be. There does seem to be clearly something psychologically interesting going on when we engage with a work of fiction. If we shouldn't explain our behaviors, emotions, beliefs, and desires in terms of a distinct attitude, then how should we?

3. Foundations of the fictional stance

We spend a great deal of our lives engaging with fictions: films, novels, TV shows, etc. So it is

important to develop an account of these engagements that appeals to commonsense ideas and mental capacities. The fictional stance provides just that. Each of the following three concepts may seem like a special skill or type of knowledge. However, I argue that they capture aspects of our experiences with fictions that are easily understood in terms of standard cognitive capacities.

3.1 Works of fiction

It is well beyond the scope of this paper to present a fully articulated ontology of fiction. Nevertheless, it will be important in what follows to understand what we mean by a work of *fiction* as opposed to nonfiction.

There are three ways (at least) to distinguish between fiction and nonfiction. First, the objects of our engagement are not actual. By this, I mean that we do not engage with a concrete object that we can encounter in our spatio-temporal world. This condition leaves open the possibility that fictional entities are abstract artifacts (Kripke 2013, Thomasson 1999, Schiffer 1996, Salmon 1996), possible objects (Lewis 1978, Plantinga 1974), definite descriptions (Russell 1905; see also Quine 1953), or even eternally existing abstracta (Priest 1997; see also Meinong 1904/1981). It also leaves open the possibility that fictional entities could exist or existed in the past.

Second, fictional entities, as well as the fictional world in which they are found, have been created by an actual person or group of people (author, playwright, filmmaker(s), etc.)—or, at least have been called into being by some person or persons. I think that the more commonsensical claim is that fictional entities (or representations or descriptions of them) are created by an author (etc.), but at this point I will grant that fictional entities may be eternally persisting abstracta for

the sake of theoretical neutrality (i.e. fictional entities may be denizens of a Platonic heaven that authors draw upon but do not, strictly speaking, create; see Wolterstorff 1980; Meinong 1904/1981).

The two previous conditions are implicit in how we treat and talk about fictional entities. If asked, I think that audiences would readily grant them. There is also an implication of the first two conditions that is worth stressing. Fictions have *dependence conditions*. They depend on other things in order to exist (or, if not exist, then in order to be represented). Fictional entities depend on creators in order to come into existence. They also depend on particular media in order to *persist*, to continue in existence. Represented objects in a painting depend on that painting in order to persist. Film characters depend on tokens of the film, literary characters on tokens of novels, dramatic characters on tokens of plays. A fiction's persistence also depends on everyday people. If every token of *Hamlet* was destroyed *Fahrenheit 451*-style, and every person who knew about Hamlet passed away or had their memory erased, then, arguably, the character Hamlet would cease to exist (again, barring the idea that fictional entities are Platonic abstracta). We think of fictions as depending on particular media and people.

In contrast, nonfictions are about actual events and entities (either present or past). The subjects of nonfictions are not created by an author, filmmaker, etc. even if the representation of them *are*. Furthermore, the subjects of non-fiction do not depend on a particular medium, creator, or audience in order to persist. I argue that we distinguish between fiction and nonfiction in these three ways. Indeed, it is part of our experience of fiction that we understand the object of our engagement to meet these conditions. We are either engaged with a fictional medium, creating one, or recollecting a previous engagement with one.

Three challenges immediately arise here. First, a reader (or viewer) may mistake a fictional

representation for a non-fiction. Second, a reader may not *know* whether a representation is fictional or non-fictional. Finally, it is unclear how the three conditions can deal with fictions that contain actual people, places, or events, such as Napoleon Bonaparte or London, England. I will return to these three questions in the final section, after completing my discussion of the fictional stance. In fact, I think that the fictional stance can handle these concerns quite easily.

The first feature of taking the fictional stance involves recognizing that the representation with which we are engaged is a fictional one. We must know that we are engaged with a *fiction* before taking the fictional stance. This does not mean that we must always consciously keep in mind that the objects with which we are engaged are fictional; the point is that this knowledge is consciously accessible. Indeed, as we will see, this knowledge permeates and influences the judgments and emotional responses that we have towards fictional entities.

3.2. Seeing fictional objects

In the opening of this chapter, I claimed that there is a sense in which audiences "see" the fictional characters Benedict and Beatrice during a stage performance of *Much Ado*. The question of fictional transformation presents us with the challenge of explaining how this is possible, if all we actually perceive are actors on a stage. A further question concerns the difference between seeing a fictional character and imagining seeing that character, or even merely seeing a *representation* of a character.

I maintain that there is a sense in which we literally see fictional entities in visual fictions (film, plays, opera, TV shows, video games, even some paintings and photographs). In the case of nonvisual fictions, we can be said to *hear* fictional entities (i.e. listening to an audiobook or radio

story) or *think about* fictional entities (as in literary fictions). I will focus on visual fictions for now and return to literary works and auditory fictions in the following subsection.

Consider William Blake's illustration of Milton's *Paradise Lost*: *Satan Watching the Endearments of Adam and Eve*. Imagine that a friend turns to you, asking: "What do you see when you look at this painting?" Setting aside any metaphorical interpretations of this question, there are still several ways you could answer. First, you see brushstrokes on a canvas. Even more reductively, you perceive splashes of color, fine lines, and a variety of shapes. Both of these answers are true, but they do not really answer your friend's question. You may respond by saying: "I see objects: two entwined figures, another flying above with a snake wrapped around him." This is also a correct answer. Finally, there is a natural sense in which you can respond that you *see Satan*—you perceive the Prince of Darkness flying over Paradise, looking longingly down on Adam and Eve.

I understand each of the above responses as characterizing different senses of 'seeing.' First, there is strictly perceptual seeing. Second, we see objects. Third, there is identification of that object as being a particular thing, or kind of thing. We can straightforwardly refer to the first sense as *perceptual*. Call the second *object seeing*. I will refer to the third as *recognitional seeing*. I will also discuss seeing a representation *as* an object. Here, 'seeing as' can require either the object seeing, recognitional seeing, or both. I will try to make this clear in what follows. The question is, which senses of 'seeing' explains our interactions with visual fictions?

Let's examine the three types of seeing in further detail. Recalling my discussion of the perception of fictional entities from the previous chapter, we might ask ourselves what it is that we *ever* perceive, in real-life *or* fictional contexts. Philosophers of perception have extensively debated this point. According to many theorists, we only perceive very basic kinds of properties,

such as shapes, colors, depth, motion, location, and illumination (e.g. Clark 2000, Brogaard 2013, Dretske 1995,Tye 1995; see Siegel 2010 for an overview). On this view, we do not perceive objects, but rather just the surface of objects (Clark 2000). Low level properties result from retinal stimulation. All other visual properties, including objecthood or object-*type*, are the result of later cognitive processing (thoughts and judgments) of this basic sensory information (O'Shaughnessy 2000). Our perception of Blake's painting only includes lines, basic shapes, and shades of green, tan, and red. Further judgments and inferences are required to see the representation *as* Satan. In contrast, other theorists hold that we do, in fact, perceive properties beyond color, shape, etc. (e.g. Bayne 2009, Peacocke 1992, Siewert 1998, Siegel 2006). We can perceive objects and higher-level natural and/or artificial kind properties (e.g. dog or chair, respectively), causal properties (e.g. seeing A cause B), and emotional properties (e.g. being scary). In this case, we may see Blake's painting as representing certain objects (people, snake, etc.) and not merely shapes, lines, or colors.

We can be neutral concerning the content of visual perception. It is an open question whether object seeing is perceptual as opposed to a cognitive judgment. In our terms, recognitional seeing may go beyond mere object seeing. Recognitional seeing requires further cognitive processing, including a judgment that the perceived object is of a particular kind. Our capacity to *recognize* part of the Blake's painting as the Angel of Light involves a mixture of perceptual and cognitive processes, such as judging or inferring that the figure at the top of the painting is *Satan*. The knowledge that the object of our perception is a fictional representation plays a role here; contextual information, including the title of the painting and a basic familiarity of *Paradise Lost*, clues us in to the fact that we see Satan, as opposed to some random figure.

Once we break down our experience of Blake's painting in terms of the above capacities,

it seems plausible that we *see* Satan. Seeing a representation as an object utilizes the same perceptual and cognitive capacities as seeing objects in our everyday lives. I must perceive and judge that an object is our neighbor's dog. Seeing an object as being a particular *kind* requires that we perceive it its properties and perhaps also judge it to be of that kind. There is nothing perceptually unusual about seeing a representation as Satan. In the fictional case, I do not merely see a *representation* of Satan; I see the representation *as* Satan.

So there is a sense in which we see an object before us when we perceive its representation. The upshot of seeing fictional entities—as opposed to imaginatively seeing them or merely seeing a representation of them—is that it can explain the immediacy of our visual experience. We do not *pretend* to, *make-believe,* or *imagine* that we see Satan; we *do* see Satan! That is not to say that we do not also see the representation as well; again, we perceive brushstrokes, lines, colors, shapes, etc., as well as the painting's canvas, frame, and wall surrounding it.[9] We never forget that the object we encounter is a representation; we are not deluded into believing that the object of our engagement is real. This will allow us to act in certain ways towards the fictional object, but not others.

The task of this subsection is to show that there is a way in which we can be said to see fictional entities and, further, that seeing fictional entities does not require imagination, pretense, or make-believe. Instead, our capacity to see a fictional entity be entirely explained in terms of standard perceptual and cognitive processes. I grant that this is not the same as perceiving an actual, physically present object. Our seeing a representation as an object lacks many of the qualities that actual seeing possesses. However, object-seeing/seeing as is an important part of normal

[9] See also the Ernst Gombrich's discussion of 'seeing as' (Gombrich 1960) and Richard Wollheim's 'seeing-in' (1980). See also Dominic Lopes (2005) for a discussion of representational seeing. Nelson Goodman (1976) also discusses these issues, and critiques Gombrich's position.

perceptual processing, both in this case and in real life. This is also not to say that we never use our imagination while engaged with visual fictions; surely we do in many cases. The point is that imagination does not play a necessary role in how we see fictional characters.

3.3. The 'is' of artistic identification

In his seminal article, "The Artworld," Danto asks us to imagine the artistic neophyte Testadura, who encounters Rauschenberg's *Bed* for the first time (Danto 1964). Testadura doesn't quite know what to do with the piece. It's actually a bed, after all, even if it is an odd one. Should he sleep on it? Why would a bed be in an art gallery? Why does it have splotches all over the comforter, and why is it so oddly shaped? Testadura looks at *Bed* and all he discerns is a bed, not an artwork. However, as Danto points out, that's really all there is (*ibid*, 575). Nevertheless, Testadura has gotten something wrong. Danto contends that understanding just how he went wrong is greatly important for understanding what makes an object an artwork, when the artwork is (physically) nothing but the object.

Danto introduces Testadura's response to Rauschenberg's *Bed* in order to motivate his art historical/theoretical view of the nature of art. To recognize that *Bed* is an artwork, and not a mere bed, we need "something the eye cannot decry—an atmosphere of artistic history, a knowledge of the history of art: an artworld" (*ibid*, 580). To recognize *Bed* as more than the bed that it is, Testadura needs knowledge of the artworld: art theory, history, artistic intention, the works role in the current art scene, etc.

I am interested in Testadura for a slightly different reason. There is an interesting similarity between Danto's 'is' and my notion of seeing represented objects. As Danto points out, we may

see a blob of paint as well as see that blob of paint *as* a person. While watching a film, we see colored, seemingly moving images and see them *as* people—fictional people. Most interestingly, at a play we see a real person and, somehow, also see that person *as* a fictional character. How do we do this? The first step is to perceive the object before us. We then see the representation as representing objects (this may involve purely perceptual or a mixture of perceptual and cognitive capacities). But this cannot be the whole story. We need to understand how we see the figure in a painting, the person in a film, and the person on stage not only as objects, but as of *fictional* objects. This is not, strictly speaking, a perceptual capacity. We do not perceive properties like "being fictional," even on theories of perceptual content that grant that we can perceive some higher-level properties. Nevertheless, we do see fictional entities. The 'seeing' here is more like what I have been calling recognitional seeing, involving a judgment about the particular object that we see. Part of this involves judging that the object is fictional; that is, that it meets the three conditions I posited in §3.1.

Let's further draw out the analogy between seeing fictional objects and the 'is' of artistic identification. Imagine standing in front of Brueghel's *Landscape with the Fall of Icarus*. You scan the painting for the doomed young man. He isn't easy to find. You finally spot him and, pointing to the painting, you proclaim: "That white dab of paint *is* Icarus." Danto notes that this "is" needs explaining:

> There is an *is* that figures prominently in statements concerning artworks which is not the *is* of either identity or predication; nor is it the *is* of existence, of identification, or some special *is* made up to serve a philosophic end. Nevertheless, it is in common usage, and is readily mastered by children. It is the sense of *is* in accordance with which a child, shown a circle and a triangle and asked which is him and which his sister, will point to the triangle saying "That is me"; or, in response to my question, the person next to me points to the man in purple and says "That one is Lear" (*ibid*, 576).

Danto calls this the "*is* of artistic identification." Mastering the 'is' of artistic identification is

required for us to understand that Rauschenberg's *Bed* is not merely a bed, that Warhol's *Brillo Box* is not merely a Brillo box, and that Duchamp's *Fountain* is not merely a urinal. It allows us to understand that these are *artworks*. In fact, recognizing *any* artwork as not merely the physical stuff of which it is composed, but also *as an artwork* requires that we master the 'is' of artistic identification. Danto argues that the 'is' of artistic identification is essentially the 'is' of metaphor, not of predication or identity (Danto 1981). When we hear that "Juliet is the sun" or "All the world is a stage," the special use of *is* in each cases invites the audience to consider the subject of the sentence as something else. Likewise, recognizing that an object is also an art object "transfigures" the object into a new, glorified status: an artwork. We now consider the object in a new way.

Danto's 'is' has garnered a fair amount of critique. For instance, it is unclear that recognizing an object as an artwork does, in fact, require application of a metaphor. I want to be neutral here, especially since I think that Danto's suggestion depends on a theory of metaphor, a determination of which is beyond the scope of the current discussion. Luckily, I think that we can distance ourselves from this debate for our present purposes. I understand the 'is' of artistic identification in terms of *representation* rather than metaphor. When I say "that is Icarus" or "that one is me" I mean that Icarus is represented *there*, by *that* white dab of paint, or that the stick drawing represents *me*. In each case, some transfigurative process occurs; I now consider the object in a way I didn't before. The same basic idea applies to Rauschenberg's *Bed*. What I once thought of as merely a (physical) bed, I now know to be an artwork. I understand now that the object represent certain artistic ideas or intentions.

I want to make the case that the 'is' of artistic identification is a special application of seeing represented objects. Awareness of a representation as a *fictional* representation requires knowledge of a particular kind. According to Danto, to think of an artwork as an artwork will in

each case require that the audience master the '*is'* of artistic identification. I want to make a similar claim about how we see objects in artworks as fictional entities and, likewise, we see representations as representations of fictional things. During every encounter with fiction we perceive whatever physically comprises the fictional representation (film, paint, physical person, or even words on a page). We also see the representation as presenting objects. Finally, applying our knowledge that the work is fictional, we recognize that the objects we see are *fictional*. My proclamation that "That white dab of paint is Icarus" not only points out where the subject is located in the painting, but also implies that I see and understand Icarus as a part of the representation (as a part of its fictional world, if you will).

When watching a production of *Much Ado*, I may point to a brown-haired actress on stage and tell my companion, "That one is Beatrice." Statements like this indicate that I have mastered a special kind of 'is.' Following Danto, let's call this the *'is' of fictional transformation*. I see the represented object as a fictional entity. I judge that the representation before me is fictional. This allows me to think and speak about the represented objects in terms of their being fictional. As I noted in chapter 1, there is an implicit fictional operator in our speech acts about fictions. When I say "I believe that Demetrius treated Helena very poorly" after watching *A Midsummer's Night Dream*, I mean "I believe that Demetrius treated Helena very poorly [in the fiction]" (see Kripke 2013 and Thomasson 1999). The same holds for our mental attitudes; the fictional operator is implicit in our beliefs, thoughts, and desires about fictional entities (see also Matravers 1991 & 2014, Neill 1993). So I may believe"that Beatrice secretly loves Benedict [in the fictional world]." The recognition that the object of our mental attitudes is fictional forms the backdrop against which these attitudes are formed.

The 'is' of fictional transformation doesn't only apply to visual representations. It also

explains how we interact with non-visual fictions, like literature. Unlike paintings, films, or photographs, you cannot point to anything in a novel and proclaim "There is Elizabeth Bennett." We would never mistake a word on a page for a fictional character. However, we still need to recognize that the object of our engagement is a fictional representation in order to get our literary experience off the ground. One needs to recognize that the words on a page are designed to be taken up and considered as a fiction. This requires utilizing the 'is' of fictional transformation; we perceive the words on a page and also recognize that those words represent fictional objects that we can directly think about and respond to. This causes us to treat the statements found in the literary work as representing a fictional story.

4. The fictional stance

The three concepts we have discussed so far—the ontology of fiction, seeing represented objects, and the 'is' of fictional transformation—form the backbone of the fictional stance. I offer the fictional stance as a general account of our psychological interactions with fictions, including how we understand objects as *fictional* and how we come to mentally respond to them. By taking the fictional stance, we recognize a work as fictional. This recognition shapes our mental states and allows us to understand and interpret the story. It also allows us to have beliefs, desires, and emotions towards fictional objects even while we know that those objects are not real. We know that the objects of fiction are not actual denizens of our world, but we think of and respond to them in similar ways notwithstanding.

In sum, by taking the fictional stance we recognize, both perceptually and cognitively, that a representation is *fictional*. This means:

1. We know that the representation meets three conditions for being a fiction: is non-actual, is created, and depends on particular objects and people in order to persist (*the ontology of fiction*).

2. We can recognize representations as of or about particular kinds of objects (*representational seeing*).

3. We recognize the represented objects as fictional objects (*the 'is' of fictional transformation*).

When we engage with a fiction we perceive objects, images, and words: the physical medium of which the fiction is constructed. We also see fictional entities. We do not pretend to see them or imagine that we do—we actually see representations as fictional objects. I see the actor who portrays Benedict *and* Benedict the fictional character. I see the physical stuff that makes up the dramatic props: the scaffolding, the papier-mâché, the painted wood comprised the furniture, even the actors' bodies. It's not that I ever stop seeing those materials. Rather, I also see them as fictional entities. I can identify each physical thing as a fictional one; that boulder *is* a rock in Italian countryside, that brazier *is* a torch on the walls of the Italian villa, that man *is* Benedict. I have argued that an extra step is needed in order for us to understand these objects to be representations of fictional entities, beyond object recognition. This step is the 'is' of fictional transformation, through which we come to understand that the object of our engagement is *fictional*.[10]

We see representations as fictional entities. We also have emotional responses towards these entities, morally judge them, desire things for them, and believe certain things about them. Importantly, nothing about the fictional stance requires that we analyze these mental attitudes as

[10] This is similar to Robert Hopkin's notion of "collapsed seeing in" according to which film audiences typically see the events represented in the film's story, but not the representation of those events (the actors, props, etc.). (Hopkins 2008).

being different in kind from ordinary mental attitudes. I contend that our mental attitudes towards fictions are standard mental states. Our mental states contain fictional *content* (they are about fictional entities), but are of the same type as typical mental states and they utilize the same cognitive mechanisms.

That is not to say that there are no differences between how we respond to fiction and real-life objects. Consider, for example, the differences in how we would behave towards fictional and real-life objects. As Katherine Thomson-Jones (2008) points out, I do not run screaming from a movie theater when I see a frightening serial killer hiding in the shadows, as I likely would if I encountered the killer in real-life. Such discrepancies in our behaviors towards real-life and fiction are typically used to motivate the DAV. Assuming functionalism about mental states (mental states are individuated, at least in part, by their functional role), then a lack of motivation to act suggests that the mental state itself is not present, or runs 'off-line' (Currie 1995; Currie & Ravenscroft 2002). This amounts to a pretend or imaginative belief in the presence of the fictional killer, but not a genuine belief. Genuine beliefs motivate action, fictional beliefs do not.

The fictional stance offers a different interpretation of these behavioral discrepancies. We *do* have ordinary mental attitudes towards fictional objects. Our mental processes are not taken off-line. Instead, behavioral discrepancies can be explained in terms of a difference in the intentional *content*. We recognize that the objects of our engagement are not literally present to us (they are non-actual). This recognition informs our responses to fictional entities. As I stated in §2, our mental attitudes towards fictions have a kind of fictional operator; we have beliefs, emotions, desires, etc. *about* a fictional character/fictional world. My belief about the serial killer on the movie screen states: "I believe that there is a serial killer lurking in the shadows [in the fictional world]." This belief does not functionally motivate a fleeing response, as it likely would

in real-life contexts. It may, however, motivate other kinds of behaviors, such as covering one's eyes, turning from the movie screen, etc.

I think that the same sort of story can explain all of our mental attitudes towards fictions, including emotions. A further worry arises here in terms of the question of fictional response. If we know that the object of our engagement is fictional—non-actual, created, dependent—then why do we have any responses to it at all? This question concerns the *appropriateness* of our mental attitudes towards fictional entities. I have two thoughts here. First, I have argued that we see representations as fictional characters. We see the actor portraying Hamlet, but we also see the fictional character Hamlet. This means that see Hamlet *as* a person. We do not stop seeing him as a person when we also acknowledge that he is a fictional character. It is likely, then, that we will interact with Hamlet as we would if he were a real person—if not physically, then at least in terms of our psychological engagement. This means that we will emotionally respond to Hamlet, judge his actions, etc.

Perhaps this is all that we need to get our psychological responses off the ground; we recognize fictional characters as the types of things that we would typically respond to in our everyday lives. This may take place subconsciously and unintentionally. Recall that object-seeing can take place before we judge whether the object is fictional or not. Maybe some emotional responses, behavioral motivation, and even judgments also occur before we make that judgment. I proposed this in chapter 2 and will flesh out this possibility in the following chapters. We also have beliefs and desires about fictional entities while, at the same time, we acknowledge that the object of our belief and desire is not real.

In sum, taking the fictional stance involves the application of several mental abilities. First, we see representations as objects. To make this the *fictional* stance, we must also realize that the

object with which we interact is fictional, that it has non-actual content, was created by someone, and depends on particular people and things in order to persist. All this occurs implicitly and naturally upon learning the conventions of fiction. We then apply the 'is' of fictional transformation which allows us to see representations as being of fictional entities. Our mental engagements with fictions rely upon our general capacity to see objects in fictional representations and respond to such objects; we see the representation of the fictional entity as an object that thinks, feels, acts—in short, as an object that we would respond to in our everyday lives. We mentally interact with fictional entities in much the same way we would real-life entities even though we acknowledge that, strictly speaking, fictional entities are not the sorts of things that possess mental states or have things happen to them.

I grant that the fictional stance is not a complete explanation of our mental responses to fictions, especially our emotional and moral responses. Simply recognizing an object may sometimes be enough to bring about a moral or emotional response, but it often won't be. The fictional stance lays the *foundation* for further explanations of why we respond to fictional objects as we do; it makes such responses possible. I will explore other aspects of our engagements with fictions in greater detail in the chapters to come.

5. Variations on the fictional stance

In my terms, taking the fictional stance is a matter of seeing and mentally responding to fictional entities while at the same time acknowledging that those entities are fictional. This is not the only way to characterize the fictional stance. There are at least two other prominent versions of the

fictional stance, or, as they call it, the *fictive* stance. In this subsection, I will compare my account stance to Lamarque and Olsen's and Wolterstorff's positions. I will also determine whether my version commits me to the DAV.

Let's begin with Lamarque and Olsen's fictive stance. These authors grant that the distinction between fiction and nonfiction is fundamental. It is of utmost importance to both creators of fiction and audiences that they are dealing with a fiction, for this knowledge shapes how they approach it and the responses they have to the objects within it (see also Currie 1990). So far, this view is quite similar to mine. But the authors go further, stating that fictions in general are defined in terms of how an audience takes the fictive stance:

> The fictive story-teller, making up a story makes and presents sentences (or propositions, i.e. sentence-meanings) for a particular kind of attention. The aim, at first approximation, is this:
>
> For the audience to make-believe (pretend or imagine) that the standard speech act commitments associated with the sentences are operative even while knowing that they are not.
>
> Attending to the sentences in this way is to adopt the *fictive stance* towards them (Lamarque and Olsen 1994, 43; quoted in Matravers 2014, 54-55).

This view conforms to what I have been calling the DAV and what Matravers calls "the consensus view." Understanding a work as fictional requires a distinct mental state, either imaginary, pretend, simulated, make-believe, or some variation of one of these. Matravers denies this requirement and also denies that there is a fundamental difference between fiction and nonfiction. In making the first claim, he seems to argue against fictive/fictional stances in general. Could it be that any version of the fictive stance requires a unique mental state? If so, that would be a serious challenge for my position.

The main difference between Lamarque and Olsen's fictive stance and my fictional stance is that my view is not primarily a theory of *fiction*, but rather how we *recognize* and *respond to*

fiction. In taking the fictional stance, we recognize that the object of our engagement has the three conditions I listed in the previous subsection: it is about something non-actual, it is created, and it has certain dependence relations. These conditions make a narrative a fictional one whether we know that it is or not. We take the fictional stance in response to our acknowledgement of a narrative meeting those conditions. So if imagination or make-believe is involved in our engagement with fictions, those aspects of our psychological engagement do not, on my view, make the narrative a fictional one. This means that there is no essential connection between imagination or make-believe and fiction. This should eliminate any worry that Matravers' might have that my fictional stance conforms to the consensus view.

Nevertheless, a worry may remain that the fictional stance commits me to some kind of DAV, as Lamarque and Olsen's does. I think that this worry can partially be eliminated by the previous point, but it is worth examining more closely. The fictional stance is not primarily a kind of mental state. It is a way of treating a particular kind of intentional content. The fictional stance does not utilize a distinct or unique kind of mental state or process. In fact, we utilize the same kinds of mental states and processes as we would for a nonfictional representation, or, indeed an everyday object with which we are confronted. Here Matravers and I part ways; he argues that our mental states towards representations are run offline. I deny this. Our mental states are run *online*, but with fictional content. Asymmetries in behavior and motivation can be explained accordingly; difference s result from the content of those processes and states. We have genuine perceptions, beliefs, desires, etc. *about* fictional objects. By "taking the fictional stance" I simply mean that we recognize the fictional nature of the object with which we are interacting.

The second version of the fictive stance that I want to consider comes from Wolterstorff's *Worlds and Works of Fiction*. Wolterstorff proposes that the fictive stance is something that an

author does when she projects (creates) a fictional world. Wolterstorff's theory of the nature of fictional worlds is quite complex and can be ignored for present purposes; suffice it to say that fictional worlds are sets of particular propositions. We can already see one main difference between this position and my own. Wolterstorff argues that *authors* take the fictive stance during acts of creation. In contrast, I take it that both authors and audiences take the fictional stance during their interactions with fictions.

Wolterstorff equates the fictive stance with a kind of linguistic force. We can treat the same state of affairs—e.g. "The king is dead"—with different moods. We can assert that the king is dead, wish that the king is dead, promise that the king is dead, ask whether the king is dead, etc. (*ibid*, 231). Creators of fiction take ordinary sentences but do not assert them, or even pretend to assert them. Instead, the creators *present* or *offer* sentences for an audience's *consideration*. This allows us to reflect on, ponder over, and wonder about the content of the fiction without pretending or make-believing that it is real:

> [The author] does this for our edification, for our delight, for our illumination, for our cathartic cleansing, and more besides. It's as if every work of fiction were prefaced with the worlds 'I hereby present that…' or 'I hereby invite you to consider that…' Of course all of us on occasion invite others to take up this stance toward some state of affairs. But most of us do so only incidentally. The novelist and the dramatist make a profession of what for the rest of us is only an incidental diversion. That is what makes them fictioneers (*ibid*, 233).

Thus, like Lamarque and Olsen, Wolterstorff contends that the essence of fiction depends on an author taking up the fictive stance, by presenting material in a certain kind of way. Presumably, the audience would know that this is what the author intended to do, and so realize that the narrative they are considering is a fictional one.

I think that understanding a fiction as a fiction, as opposed to a nonfiction, concerns the properties of the work itself rather than a mood we take toward it. Nevertheless, I agree with Wolterstorff that the fictional stance should apply to what authors to create fictions. I disagree, though, that this is the sole application of the fictional stance. This may just be a stipulation on Wolterstorff's part. He does have an account of an audience's engagement with fictions, but does not include it in the fictive stance. This also separates Wolterstorff from other philosophers, like Currie (1990) and Walton, who argue that fiction is essentially a kind of linguistic force we take towards certain representations. I think that this is the path to the dark side, for it easily allows us to posit unique attitudes for our engagements with fiction. In contrast, I would argue that authors and audiences are doing basically the same thing while they are creating or responding to fictions, respectively. They are considering, thinking about, responding to a certain kind of *content*.

Finally, Wolterstorff is at pains to show that the fictive stance applies to a variety of fictional art forms. I admire this aspect of his view. I, too, argue that the fictional stance applies to all fictions—not just literature, but also film, drama, opera, video games, and maybe even nonfactual objects represented in the visual arts. One might think that this is impossible on my view, since I have characterized the fictional stance in terms of seeing as. But the fictional stance also applies to non-visual fictions. Paramount to the fictional stance is the idea that we recognize the fictional object as fictional. I put this recognition in terms of visual perception above, but it need not be. This would be true in the case of literature. We read a literary work and take the fictional stance. We typically know that the work is a fiction and that the words on the page represent fictional entities and events. This knowledge shapes our interactions with it: we know that the objects represented in the fiction are non-actual, are created, and depend on people and its particular medium. Nevertheless, we are still able to think about the fictional objects as

representations of objects: as people, places, and things.

6. Ontological housekeeping

Let's return to the three challenges to my account of fiction that I discussed in §3.1 and explore their implications for the fictional stance.

First, it's possible that an audience mistakes a fictional representation for an actual one. For example, we can imagine a viewer watching a film that is shot as if it were a documentary, but actually is a realistic fiction, like *The Boondock Saints* example from chapter 1. In this case, I would argue that the viewer does not take the fictional stance. However, while the viewer does not know that she is reading a fictional work, the work itself still meets the three conditions, and so still counts as a work of fiction. She will treat the objects in the narrative as if they were a documentation of real-life, actual events. The reader may come to have different responses towards the representation later on if and when she finds out that she was watching a fictional story.

That was a relatively easy case for the fictional stance to explain. But there are two further cases that prove to be more of a challenge. First, there may be situations in which a reader does not know whether the narrative she is reading is fictional or real. Perhaps she is reading a well-researched crime drama and the book jacket, preface, and reviews are ambiguous concerning whether the book is based on a fictional murder or a real one. Typically one of these cues will indicate that a story is fictional, if we didn't know already. Or there could be any other number of contextual clues about the fictional status of a work: what we know about the author, how the book is marketed, what our friends, social media, newspaper, etc. have told us about it, even where we

find the book in the bookstore and how it is marketed online. If these fail, then the content of the story will often indicate that it is a work of fiction. All of these cues will be indications for us to take the fictional stance. But we can still imagine that a reader simply finds a book and neither the context nor the content of the narrative indicates whether the book is fictional or not. Does she take the fictional stance in this case?

I think she doesn't. In general, it may be best to be skeptical about the fictional status of a work if we lack any indication to treat it as such. This is also Matravers' position and, perhaps, Kendall Walton's. Matravers' follows David Davies' *fidelity constraint*: we assume that an author's desire to be faithful to how actual events took place constrains nonfictional narratives. We do not have the same assumption with fictions; we think that the author was guided by some other story-telling purpose (Davies 2007). Walton's *reality principle* states that we should treat fictional worlds like the actual one except in places where we are indicated otherwise (Walton 1990). The fidelity constraint and the reality principle both suggest that fictional worlds may be like the actual one in many respects, but differ in others. In fact, we can assume that the fictional world is like the actual world unless we are explicitly told otherwise. This may not be true for some genres, such as fantasy or science fiction, as both philosophers acknowledge, but these will probably not be ambiguous cases to begin with. But in many other cases these principles hold. If so, we should not take a narrative that seems to be very much like the actual world as fiction unless we have some good reason to do so. This means that we should not take the fictional stance towards a truly ambiguous narrative.[11]

The second difficult case concerns fictions that contain propositions, objects, places, and

[11] See Matravers 2014 for a nice list of narratives that blur the lines between fiction and nonfiction narratives. In these cases, it may be best to be skeptical about the fictional status, and to recognize the aesthetic purpose behind writing a narrative that is ambiguous between fiction and nonfiction.

characters that are a part of the actual world. Two favorite examples in the literature are Conan Doyle's Sherlock Holmes stories, which take place in London, and Tolstoy's *War & Peace*, which features Napoleon's land invasion of Russia. London is a real place and, arguably, we need to recognize it as such while we are reading the stories in order for them to make sense. The same with Napoleon; Tolstoy's philosophy of history is based on the idea that the narrative he presents is based on events the reader knows actually happened.

There are two possible ways that the fictional stance can deal with the actual objects (etc.) that occur in works of fiction. First, it may be that we take the fictional stance towards the entire work, including the objects and propositions that correspond to actual world objects and states of affairs. On this view, we take the fictional stance towards something that is true in the actual world. We treat the actual object as part of the world of fiction, as something that fits together with all the rest of the created, non-actual objects. This weaves together the (actually) true and (actually) false propositions in the story into one fictional narrative. The challenge for this view would be to explain how we see the proposition or object as fictional and *also* as true in the actual world.

A second possibility is that we only take the fictional stance towards the clearly fictional elements in the story and not the actual ones. This view has the advantage of keeping separate the (actually) true and (actually) false aspects of a narrative. It may seem like this advantage comes at the expense of a disjointed narrative. If we treat the non-actual events in a story separate from the actual ones, how do we understand them to be part of the same fictional narrative? One way to do so is to keep in mind that works of fiction are created by authors who themselves take the fictional stance. Authors create a representation of a fictional world that bears some resemblances to our own. The reader should keep in mind the real aspects of the story for aesthetic reasons. Readers compartmentalize the actual content in a narrative so as not to confuse it with the fictional content.

We draw upon our knowledge of London to inform ourselves about Watson and Holmes' antics, but we nevertheless know that the events in the fiction are mainly fictional and so non-actual. Still, both elements work together to create a fictional world that bears many resemblances to our own, perhaps according to Walton's reality principle.

I won't decide here which of these two positions is correct. The solution may boil down to whether readers actually *do* compartimentalize actual truths separately from merely fictional ones (Matravers 2014). It may be that readers do not compartimentalize the actually true information in their mental model of the fictional world. Nevertheless, they are able to tell which propositions are true of the actual world and which only of the fictional one while engaged with the fictional story. Both positions, I think, are compatible with the fictional stance. It's also worth noting that the actual people, places, and things involved in fictional stories are slightly different from their real world counterparts; they are *fictionalized versions* of actual things. Napoleon says and does slightly different things in *War and Peace* than he did in real life, different events take place in Sherlock Holmes' fictionalized London. We bring information about these characters and places to our engagement with the fiction, but then let the fiction guide our further thoughts and responses to them. This may be a point in favor of the view that we take the fictional stance towards the entire narrative, including actual objects contained in it.

Another major area of debate in the philosophical literature concerns the ontology of fictions and fictional entities. Do fictions, and the objects presented in them, actually exist? Clearly they are not concrete objects, but do they exist in some other way? There are several possible answers to this question. First, it could be that fictional entities are pretend objects; we pretend or imagine that they exist, but they actually do not. This is the view held by many theorists of make-believe, including Walton (1990; see also Sartre 1940/2004). Second, it could be that fictional

entities are not actual objects, but rather possible ones; they exist in some other possible world (Lewis 1978 & Plantinga 1974). Third, perhaps fictional entities are linguistic constructs, such as Russellian definite descriptions (Russell 1905). This is an anti-realist position; fictional names do not refer because there is no actual thing for them to refer to. We speak as though fictional entities exist, but they do not really (see also Quine 1953). Fourth, fictional entities might be eternally existing abstracta, as Meinong famously posited (1904/1981). Fictional entities are not created by authors. They also do not exist, but rather 'subsist' (a different kind of existence) in an abstract realm and are drawn upon by writers of fiction (Wolterstorff holds a similar position, 1980). Finally, fictional entities might be abstract artifacts: abstract in the sense that they are not concrete particulars, artifacts in the sense of being created (Kripke 2013, Salmon 1998, Schiffer 1996, Thomasson 1999 & 2003).

Three of these views are inconsistent with the fictional stance and one would be a challenge to incorporate. The first is the Meinongian thesis that fictional entities are abstracta that exist like Platonic forms—they are eternal, not created. The other inconsistent view is that fictional entities are possible objects. In the first chapter, I argued that fictional entities are probably not possible objects belonging to possible worlds, because possible worlds are complete and consistent, and fictional worlds need not be. That is, for any proposition A, in a possible world B, A must be either true or false. This is not so in fictional worlds. To use the famous example, the question "how many children does Lady Macbeth have?" has no answer. One can say that "Lady Macbeth has fourteen children" but this proposition is neither true nor false in the fictional world, because there is nothing in *Macbeth* to indicate an answer. The number of children that Lady Macbeth has had is indeterminate, so the fictional world is also indeterminate. If Lady Macbeth was a denizen of a possible world, then there would be an answer to this question (see also Thomasson 1999 and

Wolterstorff 1980 for similar arguments; see Priest 1997 for a counterargument). Finally, the pretense view directly appeals to distinct mental attitudes and make-believe. Perhaps there is a way to have a pretense-based view that denies any non-standard mental states for fiction, as we saw in §2, but this view may still commits us to a weaker form of the DAV according to which distinct processes and states (pretend ones) are used in our engagements with fiction.

This leaves the Russellian-type definite description view that denies existence to fictional entities and the abstract artifact view that grants it. Both theories have their virtues and both have their potential flaws. Both also seem to be compatible with our basic understanding of fictions as created by authors. I mentioned in chapter 1 that I favor the abstract artifact theory. The question is whether the fictional stance entails this view. I do not intend for it to do so. Although questions concerning the ontological status of fictions and our psychological interactions with them are related, I do not think that there is anything about the fictional stance that requires us to be a realist about fictional objects. In fact, it might not matter which of these two positions is right if one is solely concerned with the psychology of fiction. This is because both of these views deny that fictional characters possess spatiotemporal existence. All of the issues concerning our psychological interactions arise from the fact that fictional entities are not directly present to us. How we feel about the ontology of fiction may not matter for our psychology of fiction, at least if we accept the descriptivist or the artifactualist theories, both of which seem compatible with the fictional stance and how we tend to think of and interact with fictional entities.

So the antirealist position is compatible with the fictional stance, at least in principle. In practice, however, most antirealists appeal to some version of pretense in order to explain our mental attitudes and speech acts about fictions due to issues concerning reference to nonexistent entities (I will return to these momentarily). This gives us some reason to be wary of joining with

the antirealists.

Let's examine Thomasson's realist artifactual theory more closely to see if it fares better. This view states that fictional entities are abstract artifacts, objects which have no spatiotemporal location (and, so, are *abstract*). They exist, but as neither as objects of make-believe, *a la* Walton, eternal abstracta according to Meinongians, nor as real (but non-actual) citizens of another possible world. Fictional entities are not alone in being abstract artifacts. Thomasson lists everyday objects such as theories, laws, governments and literary works as similar constructed abstracta (see also Kripke 2013). Like fictional characters, these objects are "tethered to the everyday world around us" by their relation to their creators and copies of their instantiation onto a physical object (e.g. a written down or spoken law).

Fictional characters can be found "in" literary works, but here a similar question arises—where *are* literary works, if anywhere? We possess copies of *Emma*, but the work itself does not exist in a particular location and the character Emma exist in the location of a copy of the novel (Thomasson 1999, 37). Furthermore, we must treat fictional characters as *abstract* artifacts because they cannot be located in the locations or time they are purportedly supposed to be. We would be disappointed if we were to drive the thirty some-odd minutes south from London to the manor Hartfield in Surrey and hope to find Emma Woodhouse waiting inside. We would not find Emma there and we would also have committed a strange categorical mistake. Competent readers understand this as they engage with fictions. As Thomasson states: "[i]n our everyday discussion of literature we treat fictional characters as created entities brought into existence at a certain time through the acts of an author" (*ibid*, 16). The sense in which fictional characters *do* seem to be real stems from their created nature: "We do not describe authors as discovering their characters or selecting them from an ever-present set of abstract, nonexistent or possible objects. Instead, we

describe authors as inventing their characters, making them up, or creating them, so that before being written about by an author, there is no fictional object….a work of fiction is necessarily tied to its particular origin" (*ibid*, 6-7).

The abstract artifact view is one possible way of understanding the ontology of fictional characters, a view that is compatible with the fictional stance and the SAV. There are two main issues worth addressing here: whether sentences in fictions possess truth values and, relatedly, reference to fictional entities. We already saw that one reading of the fictional stance states that some propositions in a fiction do possess a truth value—that is, they can be true or false about the actual world. These are sentences like "Napoleon and his army invaded Russia" and "Baker Street is in London." Other sentences do not possess a truth value for the actual world, but rather for the world of the fiction. They can be true or false *in* or *according to* the story. This is the reading that I am inclined towards because it seems to fit nicely with the SAV. But it is not the only possibility. Again, I do not think that there is an essential relation between the view that our mental states towards fictions are genuine and a positive answer to the question of whether fictions have truth values and authors make genuine assertions. The alternative might be that authors *pretend* to make assertions when constructing a fiction. But considering my general anti-pretense position, I am disinclined to take this route.

Here's another way in which we can understand the semantic content of a fiction. Kripke (2013) and Thomasson (1999), amongst others, have each argued that there are two general kinds of sentences we can construct about fictional entities: internal and external sentences. Some sentences possess content that is internal to the fictional world. "Claudius killed his own brother," "King Joffrey was a tyrant," and "Frodo carried the Ring to Mordor" are all sentences internal to a fictional world (*Hamlet*, *A Song of Ice and Fire*, and *The Lord of the Rings*, respectively). Other

sentences possess content about the actual world, a world external to the world of the fiction. "Sherlock Holmes was created by Sir Arthur Conan Doyle" and "Emma Woodhouse does not exist in this world"[12] are both external sentences, sentences about fictional entities (or definite descriptions) but not about the world of the fiction. External sentences are genuine propositions that are actually asserted and possess a truth value (both true, in this case). They also possess names that refer to fictional entities; on both of these philosopher's views, fictional entities are abstract artifacts. "Frodo" refers to the abstract artifact Frodo, "Claudius" to Claudius the fictional character, etc.

Internal sentences are trickier. Kripke argues that the internal sentences possess the implicit, unspoken fictional operator "in the story." This means that when we say "Claudius killed his own brother" what we really mean is "Claudius killed his own brother [in the fictional world of *Hamlet*]". Kripke does not think that these are genuine assertions (see Kripke 2013, lecture II). Instead, he thinks that the speaker *pretends* to make assertions about states of affairs internal to a fictional world. Only external sentences are genuine assertions. We also *pretend* to refer to a fictional character. This is because fictional entities, if they exist at all, are not the types of things that can kill, think, etc. They are abstract artifacts, which cannot possess the right kind of properties (having a body, being about to think, etc.)

I want to resist the pretense move if at all possible. Unfortunately, most realist philosophers have adopted some kind of pretense to explain certain types of sentences about fictions (including Thomasson, Searle, & Schiffer; Salmon 1998 being an exception).

Why propose that fictional entities exist and can be referred to for *some* sentences, but not

[12] This is a negative existential, a sentence type that has plagued philosophers for decades. See Thomasson 2003 and Kripke 2013 for a discussion.

all of them? As Nathan Salmon says, "it is like buying a luxurious Italian sports car only to leave it garaged" (Salmon 1998, 298). This means that one would have to propose a different way to make genuine assertions about fictions and refer to fictional entities. I am inclined towards a view that we refer to actual, abstract artifacts in both external sentences *and* internal sentences. The main challenge for this view is to explain how we speak about fictional entities as possessing properties that they don't seem to actually possess. Abstract artifacts cannot possess properties like "being rich" or "living in London" because they are not concrete objects. Only concrete objects can have these kinds of properties, because only concrete objects can live in London and have money (etc.).

This is an area that is fraught with controversy. As Kripke notes, we need to be able to say that a student is speaking the truth when she says that "Emma Woodhouse is rich" (even if she is pretending) and false when she says that "Emma Woodhouse is poor." One possibility is that the abstract artifact really does possess the property of "being rich." But this stretches the notion of an abstract entity; how can something that doesn't take up space have a lot of money? Another, better explanation stems from the fictional stance itself. I agree with Kripke and Thomasson that statements made in internal fictional contexts—sentences that take the fictional world as their content—should be understood as implicitly describing what is true to the story. Hence the notion of a fictional operator. When I say that "Emma Woodhouse is rich," my statement is true according to the story but false in the actual world.

So far this agrees with most realist philosophers' theories of reference and predication. But where other realists propose pretense in order to explain our mental attitude during these statements, I deny this. When I make statements like "Emma is rich" I am taking the fictional stance: I represent this object as having properties that I know that it does not really, physically

possess. Just like speaking about lightning as desiring to strike in a certain location is, strictly speaking, a categorical mistake, so too is it a mistake to say that Emma is rich if we do not include the fictional operator (see Dennett 1997). Neither speech act necessarily requires pretense. When I say that the lightning wants to strike the metal pole, I do not pretend that the lightning possesses mental states. I am treating it as if it actually does possess those states—and, more than that, I am accepting that it does have that desire. A scientifically-minded friend might come along and say, "Well the lightning doesn't *really* want to hit the pole. Lightning can't *want* anything, after all." The likely response would be that this was "a manner of speaking." I talk about the lightning as if it desires certain outcomes, but I know that it doesn't really. Still, I speak and think about the lightning as if it really does possess mental states.

The same may apply to our speech acts about fiction. I do not pretend that the fictional character Emma Woodhouse is rich. I really do think that this character is rich *in the context of the story*. In the fictional story, Emma Woodhouse has all sorts of properties. By taking the fictional stance I treat the representation of the fictional character in some ways as if Emma is an actual person. I do not do this by pretending that Emma is real, since I simultaneously treat her as a fictional entity, but rather in terms of how I speak, think, and respond to the character. Moreover, when I speak about Emma Woodhouse, I refer to Emma the fictional character—the character that actually lacks properties concerning wealth. But I still refer to that character.

My claims concerning the ontology and semantics of the fictional stance need further explanation and elucidation. It is well beyond the limits of this dissertation to do so. I leave it as an exciting and philosophically rich project to develop an ontology and semantics of fiction that is fully compatible with the fictional stance. For the rest of this dissertation I will explore the psychological side of our interactions with fictions, now that we have established how we come to

see fictions as fictions and how it is possible to have psychological responses to them. In the next three chapters, I will explore different ways in which we psychologically participate with fictions: attributing mental states to fictional entities (chapter 4), our emotional responses to fictions (chapter 5) and our moral judgments of fictional characters (chapter 6). The fictional stance is the foundation for each of these chapters, as well as the following chapters that dissolve the sympathy for the devil phenomenon, the puzzle of imaginative resistance, and the question of moral learning.

Chapter 4: Understanding fictional characters

1. A fictional mind-meld

Navigating our social climate would be so much simpler if we could be a little more like the Vulcans. *Star Trek*'s Mr. Spock simply places his fingers against someone's temples and instantly accesses her thoughts, memories, feelings, and past experiences. This allows the Vulcan to immediately understand (or, at least, become familiar with) another's experiences and mental states in a deeply personal, intimate way.

Unfortunately, we are not Vulcans. We can't simply mind-meld with another person to learn what she is thinking and what she has been through. But that does not mean that we mere humans are unable to access the mental lives of others. It is often relatively easy to tell what another person is thinking. We attribute mental states to others on the basis of their behaviors, verbal reports, facial expressions, body language, and even bodily reactions like perspiration, muscular tension, etc. For example, I guess that my friend is thirsty if she goes to the kitchen to get a glass from the cupboard. I also guess that she believes that the third cupboard from the left contains glasses if that's the one she reaches toward. I predict that my friend will then go to the sink and turn on the tap. This requires that I attribute beliefs, desires, and intentions to my friend.

We can also attribute *emotions* to others. You watch as your sister gets increasingly irritated and angry with her colleague's insensitive comments about a coworker. She crosses her arms, frowns, and scoffs at her colleague's remarks. You may watch as your friend becomes frightened and anxious as he watches a Hitchcock thriller, as he grabs the sides of his chair or clutches his

face in nervous anticipation, eyes wide. Indeed, we are generally quite good at telling when someone feels irritated, scared, sad, etc. The same goes for other mental states; we often have little trouble attributing both mental state *types* and *content* based on another's facial expressions, bodily responses, and behaviors.

In fact, we seem to have an advantage over the Vulcans when it comes to certain kinds of *social cognition*, our ability to attribute mental states to other people (aka, *mindreading*) and to predict their actions. It is unclear whether Mr. Spock can really understand those around him (at least, the non-Vulcan around him!). This is obviously true in the *Star Trek* series, when the sayings and antics of Captain Kirk and the rest of the Enterprise crew continuously puzzle and annoy him.

This chapter explores our human capacity to understand the mental states of fictional entities. While watching a film or reading a novel, we want to know what a character thinks, feelings, and believes. This isn't always explicitly told to us by a narrator or voiceover, so we have to use our own general social cognitive abilities to glean a character's thoughts, beliefs, emotions, and desires. This understanding takes the shape of *propositional* "knowledge that" (for instance, knowledge that another person is sad) as well as *experiential* knowledge (I have the same kind of sorrowful feeling as another and so have a similar, but not identical, emotional experience).

The proponent of the DAV has a ready explanation of our social cognition of fictional characters. We encountered it back in the first chapter: *simulation theory* (ST). Consider one example from the opening of this chapter: we watch our sister becoming annoyed with her colleague who is spreading rumors about another coworker. How do we know that our sister is annoyed, and what she is annoyed about? Simulation theorists argue that we imagine (roughly) what it would be like to be in our sister's situation: her beliefs, thoughts, emotions, perspective. Doing so allows us to understand how *we* would feel if we were "in our sister's shoes." From there,

we judge that our sister must feel the same way: annoyed about her colleague's callousness.

This is a rough sketch of how we simulate more generally; we imagine ourselves in a target's position, imaginatively adopting their mental states, and judging that they must feel and think as we do while engaged in the imagining. ST also applies to fictional characters. We can imagine ourselves in their situation just like a real person's. Moreover, we can see how ST is compatible with the DAV. ST proposes distinct mental processing during our social cognition of both actual and nonactual targets. So we may have imaginative beliefs, imaginative emotions, and imaginative desires about our target, but not stereotypical ones. Our simulated mental states are supposedly run offline, disconnected from their typical functional role. Indeed, ST has become the dominant theory of mindreading in contemporary aesthetics (see Currie & Ravenscroft 2002, Freedberg & Gallese 2007, Goldman 2006, Harris 2000, Nichols 2004, and Walton 2006).

We will explore the question of whether ST is committed to a distinct cognitive attitude in §3. If so, and an empathy-based ST is the best explanation of our social cognitive capacities for fictional entities, then the SAV is in trouble. The SAV is committed to the idea that we utilize non-imaginary, stereotypical mental states during our engagements with fiction. This is incompatible with the basic principles of ST.

There is hope for the SAV. In this chapter, I will reject ST as the best theory of how we understand the mental states of fictional entities. In fact, ST faces several challenges as a theory of social cognition in general: it doesn't seem to be able to explain the breadth of our social cognitive capacities and, on many versions of the theory, it leads us to an undesirable conclusion concerning our ability to introspect our own mental states. Even if ST could surmount these difficulties, it faces challenges when applied to fiction, as we will see.

I will propose a different theory of how we understand the mental states of fictional entities:

a modified version of *theory-theory* (TT). Philosophers of art do not typically appeal to TT in order to explain our social cognition of fictional entities. This is likely because, unlike ST, TT does not appeal to imagination and imaginary mental states. While TT does not appeal to the consensus DAV, it is perfect for a proponent of the SAV. TT states that we attribute mental states to others through a process of inference-drawing from tacit folk psychological knowledge about mental states and particular perceptions and judgments about a target's behaviors and bodily expressions, to conclude that the target must think or feel a certain way. TT faces several standard objections in the social cognition literature, most notably that it proposes an overly complex cognitive architecture. Perhaps we can directly perceive the mental states of others, as opposed to inference or simulation. *Direct perception theorists* (DPT) argue just that. It certainly feels like we can directly perceive when another person is angry or sad, or even that she has certain intentions or desires. TT also faces the challenge of explaining the immediacy of or social cognitive capacities.

I will supplement the traditional account of TT with an account of *social referencing*: a heuristic model of how we quickly understand our social surroundings (see Bermudez 2005). I argue that this updated version of TT can adequately account for the challenges the standard account faces. It will also be a solid foundation for explaining how we understand the mental states of fictional entities in a way that is completely compatible with the SAV. We use stereotypical mental states and mechanisms when we theorize about fictional minds.

In §2, I will explore a scene from Steve McQueen's *12 Years a Slave* and present the basic explananda for a theory of how we understand fictional characters' mental states. I will then consider ST more thoroughly and how it applies to our engagements with fictions before critiquing it. §4 discusses two alternatives to ST: DPT and TT. I present challenges to both views before moving on to my modified TT and social referencing. I conclude by showing how my version of

TT can best explain our social cognition of fictional entities.

Social cognition has its own extensive literature in philosophy, psychology, and neuroscience, with implications ranging from childhood development and autism, basic cognition, language, evolution, and animal cognition. It is impossible for me to do justice to the extent and intricacies of the debate in this chapter. Nevertheless, I think that my arguments in this chapter are enough to show the extent and importance of mindreading for our understanding of fictions. We can also make some headway in explaining an extremely significant aspect of our social lives in general.

2. A case study in mindreading

Let's consider a scene that compels its audience to mindread a character's mental states. I will use this scene as an explananda for an adequate theory of fictional social cognition throughout this chapter. If a theory cannot explain how and why we engage with this character, then we should look elsewhere for a theory that can.

The scene I have chosen occurs late in Steve McQueen's 2013 film, *12 Years a Slave*. Solomon Northrup (played by Chiwetel Ejiofor), is a free black man living in New York with his family. He was tricked, captured, and forced into slavery in the Deep South. Northrup is shipped from plantation to plantation and master to master and eventually finds himself picking cotton on a plantation owned by the cruel, violent, drunken Edwin Epps (Michael Fassbender). Northrup has cautiously formed friendships with several of the other slaves, including the pretty, innocent, hard-working Patsey (Lupita Nyong'o). Patsey has already asked Northrup to end her life after Epps

repeatedly rapes and beats her. Northrup refuses. Northrup's constant fear and anxiety reaches its zenith when a white man named Treach comes to work on the estate. After the two men converse, Northrup begs Treach to bring a letter to his friends in the north, men who can vouch for Northrup's freedom. Treach agrees to take the letter. Sadly, Northrup later discovers that Treach did not deliver the letter, but rather disclosed the scheme to Epps—no doubt to raise his own favor and gain some easy money in the process. Northrup cleverly talks his way out of trouble when Epps confronts him, but the damage is done.

This is Northrup's bleakest hour. Northrup quietly, resignedly continues his work on the plantation, not outwardly communicating his despair. But at the funeral of a fellow slave, Edward, Northrup finally breaks down. A fellow slave woman sings "Roll, Jordan, Roll" over the burial site. The other slaves' faces are passive, seemingly resigned to their fate on the plantation. But as the camera focuses on Northrup, we see another story entirely. A complex emotional narrative is expressed in his face. The other slaves sing in the background, but Northrup does not join in. Instead, we see a look of bewilderment pass over him, as if he doesn't quite understand what has happened or where he is. There is some shock there as well, perhaps from Treach's sudden betrayal and his own rapid transition from hope to despair. There is also sorrow, but at first it is rather subdued. Finally, Northrup joins the others in song. Only then does he break down, quickly expressing grief and even anger as he looks skyward, perhaps appealing to a seemingly-absent God. The scene ends with Northrup's voice rising above the other singers', his face expressing the deep hurt that arises from the hopelessness of his situation.

It is impressive that a relatively simple scene—one long, close-up shot of a singing man's face—is able to communicate so much and so effectively. How am I able to understand so much about Northrup's mindset from this brief shot, one in which no linguistic communication occurs?

Context helps; I am quite familiar with Northrup by this point in the story. I know where he has been and what obstacles he has faced. I have developed a strong bond with him, a bond that this scene both relies on and reinforces.

What's interesting is how *easy* it feels to glean Northrup's current state. I must pay attention to the story as it unfolds, as well as Northrup's face throughout the scene. However, I do not seem to engage in a great deal of deliberation, thinking through and weighing potential options about what Northrup might be thinking at that time—at least, not consciously. Instead, I make quick, fairly accurate guesses about Northrup's thoughts and mood. On some occasions I might have to consciously and deliberately consider a target's facial expression, behaviors, etc. in order to appropriately attribute mental states to her. I may even have to put myself in her shoes, taking her perspective and imagining that it is my own. However, it often seems like we can immediately tell what someone thinks and feels, as I could with Northrup. How is this possible?

One way to access a person's mental state—and especially emotions—is by examining her face. This scene in particular would be far less effective at communicating Northrup's state of mind if the audience viewed a full-body shot instead of a shot framing his face. There is something about the human face that effectively communicates and expresses how one currently feels. Alvin Plantinga (1999) calls the face of a film character a "scene of empathy"; a direct, basic source of our understanding and feeling with her. That certainly seems to be the case here. Indeed, this scene is extremely emotionally powerful the experience is *for the viewer*. If others are like me, then watching Northrup will have elicited a range of strong emotions, such as sorrow and compassion for Northrup and anger and betrayal on his behalf.

I take our responses to this scene as data concerning how we understand fictional characters' mental states. There are several points in need of explanation. First, we seem to have

propositional knowledge of Northrup's mental states automatically, without conscious deliberation or inference-drawing. I just know that he feels sorrow and believes that he has been betrayed. There is a sense of ease and immediacy to our ability to know how Northrup thinks and feels, and why.

It could be that our quick, seemingly automatic understanding of a target's mental states is caused by equally quick, automatic neural processes. These include mirroring, mimicry, and direct association (see Hoffman 2008). *Mimicry* is an "innate, involuntary, isomorphic response to another's expression of emotion" (*ibid*, 441). There are two steps involved in mimicking the emotion of a target. First, there is an automatic change in the subject's facial expression, voice, and posture at the same time as a corresponding change in the target's facial expression, vocal intonation, posture, etc. These changes then trigger the same feelings in the target as those present in the target. There is also some evidence that subjects can mirror the motor intentions of a target; witnessing movement in another (or even in a statue!) results in the subject's motor cortex to be activated, as if *she* were the one moving (Freedberg & Gallese 2007).

Mirror neurons may be the neural basis of mimicry. These neurons are triggered when one person observes the actions or emotional expression of another. This results in the same kind of neural pattern in the subject as if she were actually performing the observed action or having the same emotion herself (*ibid*, 441; see also Freedberg & Gallese 2007, Clay & Iacoboni 2011, and Decety & Meltzhoff, Goldman 2006).

Direct association occurs when we perceive a target who undergoes an event that is similar to one that we have experienced in the past (*ibid*, 441). For example, a friend of mine has recently lost her pet dog and displays sorrow-related signals (crying, having a "long-face," slouched posture, etc.). My memory of a similar situation in which I lost my pet parrot makes me also

express these signals and even consciously experience sadness. I make this association unconsciously, without thinking about it or planning to do so.

Mimicry, mirroring, and direct association might all serve as the basis of how we quickly recognize and know about another's emotional state. The perception of another person causes us to have our own emotional responses. We then attribute particular emotions to the target on the basis of our own feelings.

While some of our social cognitive abilities seem to occur automatically, there are other cases in which understanding a target's mental state requires a slower, more thoughtful and deliberate processes. Sometimes we may need to take on a character's perspective in order to know how she feels or what she will do next. *Perspective-taking* seems especially important in ambiguous scenes in which we do not know enough about a person or her context, and so we have difficulty in judging how she feels, believes, or desires.

Sometimes perspective-taking is equated with *empathy*. When a subject *X* empathizes with a target, *Y*, she imaginatively takes on *Y*'s mental states as closely as possible. *X* shares in *Y*'s mental state and, further, that *X*'s responses are *caused by* and involve the same *type* of state as Y's. In taking another's perspective, I "put myself in her shoes," and imagine what she must think or feel in a particular situation.

Like mimicry, empathy can also engender similar emotional states in the subject. Martin Hoffman (2008) lists three types of perspective-taking. The first is *self-focused*; imagining that the subject's situation is happening to ourselves evokes an empathetic response in us. For example, I think about how my friend Sarah must feel after she lost her pet dog. Putting myself in her position also makes *me* sad, as if I were the one whose dog has gone missing. Next, perspective taking can be *other-focused*, during which we attend to the victim's feelings, behavior and "current life

condition" (*ibid*, 442). In such cases, the observer's empathy may feel more cognitive than affective. I may observe Sarah's sorrow and imagine how she feels without (consciously) feeling the sorrow to the same extent, or imagining that I have gone through the same experience. Hoffman argues that self-focused perspective-taking is more intense than other-focused perspective taking, perhaps because it is more likely to lead to direct association from past events in our own lives (*ibid*, 442). Finally, perspective taking can focus on both *the self and the other* person, a mixture of the previous two types in which we focus on both how we would feel in a situation and how the target must feel.

Perhaps we have to take on Solomon's perspective in order to fully appreciate how he feels and what he has been through. Indeed, empathetic perspective taking is also sometimes equated with *simulation*. In both cases, we imagine ourselves in the situation of a target (Goldman 2006). It would be a major win for the DAV if perspective-taking requires imagination and, further, perspective taking of this kind is required for understanding fictional characters. I will reject the identification of perspective taking, empathy, and simulation in the following section—importantly, by denying that perspective taking requires imagining. My challenge, then, is to explain how we can understand difficult and ambiguous fictional scenes without appealing to imaginative processes. It could be that some other mental process—such as association, theorizing, and inference-drawing—can explain our conscious, deliberate thinking about a fictional character's perspective.

The scene from *12 Years a Slave* captures the intricacies of our relationships with fictional entities: we attribute mental states to them, attempt to understand their past and current context, and try to determine their prospects for the future. We also seem to have a direct connection to characters through our perceptual capacities. I will repeatedly return to this scene as a test case for

the main theories of social cognition, as well as my own. It may be that one of the theories is particularly effective at explaining mirroring and direct association, but not deliberate perspective taking. Ideally, though, we can discover a theory that is capable of explaining all of them together.

For ease of reference, I will call the seemingly automatic processes (mimickry, mirroring, direct association) "low level mindreading" and the slower, deliberate processes "high level mindreading." I will now turn to ST and determine whether it can adequately explain how we understand fictional characters.

3. Simulating fictional characters

3.1. Varieties of simulation

Simulation theory arose in the late 1980's with the work of philosophers like Alvin Goldman (1989, 1993 & 2006), Robert Gordon (1986 & 1996), and Jane Heal (1996). ST holds that we utilize our own perceptual, emotional, and cognitive mechanisms while mindreading. We project or imagine ourselves in the situation/context of the observed person. These mental mechanisms are run "offline," disconnected from their typical functional output (see Currie & Ravenscroft 2002). This results in imaginative mental states that are distinct from stereotypical mental states.

ST has become immensely popular in aesthetics, as well as the social cognitive literature in general. This is partly due to the breadth of ST's explanatory power. ST seems to explain child development and autism spectrum disorder (ASD) (Harris 2000, Currie 1996, & Goldman 2006). It also may be able to handle questions concerning high and low level social cognitive capacities that I discussed in the previous section. Finally, in the aesthetics literature, many proponents of the DAV adopt some version of ST (see Nichols 2006, Currie 1995, Goldman 2006, Walton 1997).

We utilize pretend input (non-genuine beliefs, for example) and garner pretend output (a pretend belief about what a character will do next) when we simulate the mental states of a fictional entity.

There are many different varieties of ST. I will briefly discuss two: Alvin Goldman's *enactive imagining* and Robert Gordon's *radical simulationism* (Goldman 1993 & 2006, Gordon 1987 & 1996; see also Harris 2000, Nichols & Stich 2000, Heal 1996, Currie 1995, Currie & Ravenscroft 2002, Walton 1997, etc.). Central amongst these theories is the idea that simulation is the key to understanding mental concepts, mental development, and social interaction.

Let's first consider Goldman's theory of enactive imagining (1993 & 2006). Goldman argues that we attribute mental states to another after we recognize our *own* mental states under actual or imagined conditions. I do not pretend to be the target person; rather, I *transform* myself, imaginatively, into her. I imaginatively take on her relevant beliefs, desires, emotions, and perspective. This is called *enactive imagining*—or, e-imagining, for short—because I utilize my own mental processes for simulative imagining. Imagining the mental states of another involves enacting certain perspectives and states, especially when we deliberately take another's perspective. Once I e-imagine myself as the target, I can introspect what I feel during this particular situation. I then judge that the target feels the same way. So there is some amount of folk psychological knowledge and inference drawing done in the service of simulation. However, such knowledge could not be developed without first engaging in imaginative, simulative capacities (compare this interpretation of the role of folk knowledge and theorizing to TT, below).

There is a further question about how the simulator infers the target's mental states. Goldman (1993) argues that the recognition of one's own mental states requires introspective access to our own thoughts, emotions, beliefs, judgments, etc. One implication of this position is that all mental states must have some consciously introspectible property. Indeed, Goldman argues

that we introspect the conscious, *qualitative feel* of each mental state. Just as the taste of chocolate has a particular conscious taste, and the color red has a particular qualitative perceptual appearance, so too for all of our mental states including beliefs and intentions. If this is true, then we can recognize our own e-imagined mental state by its conscious qualitative features.

This is a controversial position, as we will see in the following subsection. Some versions of ST seem to rely on a subject's clear and reliable introspective access to her own mental states. Without this access we wouldn't be able to understand what another person thinks or feels based on our own simulated feelings (see Carruthers 1996 for a critique).

Not all ST's require conscious introspection. Gordon (1996) offers his *radical simulation theory* as an alternative to such theories. On Gordon's view, the subject projects herself into the situation of another person, makes the relevant adjustments for differences between herself and the other person, and then sees how she would react. This sounds quite similar to Goldman's position, but there are several important differences. First, Goldman's view does not require that the simulator possess knowledge of mental concepts, such as beliefs and desires. This is important, since very young children can take the perceptive of others without knowledge of mental states (Nichols 2004a). Second, the view does not appeal to conscious introspection of our own mental states. Gordon takes these two points as benefits of his view, since it does not align him with view of self-knowledge according to which our knowledge of our own mental states is incorrigible (we cannot be wrong about what state we are in) and transparent (we always know what type of mental state we are in while we are in it) that may plague introspective accounts like Goldman's (Moran 2001).

How can we simulate without the recognition of conscious qualitative states? Gordon suggests that simulation utilizes "ascent routines" from one higher, introspective semantic level to

another, non-introspective level. Let's say that I am trying to determine whether my friend thinks that the film that we are watching is any good. I imaginatively transform myself into her state, taking on her perspective and accounting for relevant differences between us, such as her lack of expertise in this film genre compared to my relative abundance, as well as her current emotional state, energy level, etc. Now that I have transformed myself into my friend, I simply have to ask myself whether *I* think that the film is any good. I don't have to ask whether I have a particular *belief* about the film—or, more accurately, that I have formed a particular aesthetic judgment of it. There is no higher-order mentalizing at work here, no introspection at all. Rather, I ask a direct question about the *world*—in this case, about the film. Is the film any good? Yeah, it's good for what it is (a zombie apocalypse flick, let's say). I then remove myself from the ascent routine and conclude that this must be what my friend thinks as well. Note that I have attributed a particular *type* of mental state as well as its *content*; my friend has made a positive aesthetic judgment about the zombie film.

Goldman and Gordon—as well as other simulationists, like Nichols & Stich (2000), Harris (2000), and Currie & Ravenscroft (2002)—both adopt a "hot" simulative methodology. This means that we utilize our own mental state processes while attributing mental states to others. It draws upon our own capacity to form beliefs, desires, emotions, and intentions while imagining how another person feels. The important question for us is whether the imaginative state involved in simulating different in kind from typical, non-simulative states. In other words, does ST imply the DAV?

There are two places where simulation may involve a distinct attitude type: the mental states used in *imagining* the target, and the *resulting* mental states. Currie (1995) and Walton (1990) both argue that the imaginative state involved in simulation *and* the resulting mental states

are of distinct in kind. This, as we've seen, is due to the fact that they are run offline and possess different functional roles than stereotypical states. Goldman (2006), in contrast, argues that the e-imagined states are not stereotypical beliefs, desires, etc.—also for functionalist reasons—but that the mental state that results from e-imagining *is*. We have a genuine belief that our friend feels a particular way or makes a particular kind of judgment. Finally, still other theorists distinguish between different kinds of mental states. Currie & Ravenscroft (2002) deny that the beliefs, desires, intentions, and judgments involved in simulation (at both stages) are genuine. But emotions *are*. These authors adopt a feeling theory of emotions according to which emotions are constituted by affective feelings. Affective feelings are certainly present during simulation, so the emotion is a genuine one.

So it seems generally accepted that some distinct kind of mental state goes into or is the result of the simulative process. This makes ST a good psychological basis for the DAV because it explains our understanding of fictional entities in terms of the imagination. Indeed, many writers on the psychology of fiction either explicitly (Feagin 2011, Freedberg & Gallese 2007, Goldman 2006, Harris 2000, Nichols 2004b, Walton 1997, etc.) or implicitly (e.g. Kieran 2010) buy into some version of ST. This makes sense given the assumption that imaginative processes require that we run our standard mental states and processes offline, and, further, that *this* requires simulation.

I will now examine Gregory Currie's prominent simulation-based psychology of fiction. In much of his early work on fiction, Currie (1990 & especially 1995) adopts a version of the ST that holds that mental processes must be run offline, disconnected from their typical motor/behavioral output and, further, that offline mental states are of a different type from online mental states. Currie has backed away from a strong simulative approach in some of his recent

work, but the general assumption of offline ST remains the same (see Currie & Ravenscroft 2002). Like Walton, Currie argues that our basic psychological interactions with fictions involve games of make-believe, games which are based on imaginatively simulating fictional actions and the minds of fictional characters:

> I take on, temporarily, the beliefs and desires I assume someone in that situation would start off by having; they become, temporarily my own beliefs and desires. Being, thus temporarily, my own, they work their own effects on my mental economy, having the sorts of impacts on how I feel and what I decide to do that my ordinary, real beliefs and desires have. Thus I might start off my imagining by taking on in this way the beliefs and desires and also the perceptions of someone who suddenly sees a fierce and enraged lion rushing towards him. These beliefs and desires then operate on me through their own natural powers; I start to feel the visceral sensations of fear, and I decide to flee. But I don't flee; these beliefs and desires—let us call them pretend or imaginary beliefs and desires—differ from my own real beliefs and desires not just in being temporary and cancellable. They are also, unlike my real beliefs and desires, run 'offline,' disconnected from their normal perceptual inputs and behavioral outputs…The function of the simulation is not to save me from a lion, since I am not actually threatened by one, but to help me understand the mental processes of someone who is (*ibid* 252, 253).

Here Currie presents a standard simulative account: we imaginatively take on the mental states of another as closely as possible, and run our "mental economy" in order to see how we would feel in that situation. This makes Currie's view more like Goldman's e-imagining, rather than Gordon's radical simulation. Note, also, how Currie explicitly adopts the DAV; our mental attitudes about imaginings are not stereotypical mental attitudes, but rather pretend/imaginary ones. This is because these states lack their typical functional role.

Currie's account depends on two further features of simulation: imagination and identification. According to Currie's ST, audiences are intended to adopt imaginative attitudes toward fictional characters and situations (Currie, 1990). Currie distinguishes between *primary imagining* and *secondary imagining*. Primary imaginings involve what is true in a story, "those things which it makes fictional" (1995, 255). We adopt the fictional beliefs necessary to maintain

a coherent fiction, while disregarding those which contradict it. I disregard my real belief that no objects can move faster than the speed of light while I am watch *Star Wars.* I also adopt the imaginative belief that there is a supernatural Force that guides all things. When I watch *12 Years a Slave* and come to learn that Treach has betrayed Northrup, I do not acquire a new belief that "Treach has betrayed Northrup." Instead, I acquire a "belief-like imagining" about Treach's betrayal. I may also imaginatively desire that Northrup's letter plot succeeds. My imaginings run offline in each case. They are simulations of the real mental processes that take place when I acquire beliefs and desires about the real world (*ibid*, 256).

Currie's ST plays a more prominent role in our secondary imaginings about fictions. Secondary imagining is a form of simulation that occurs when we engage in an empathetic reenactment of a character's situation (*ibid*, 256). First, I put myself into a fictional character's position: I imagine what it would be like to be Solomon Northrup during the funeral scene, singing amongst the other slaves. I imagine that I, too, have been wrongfully enslaved and taken away from my family for years, beaten, over-worked, and finally betrayed by someone I trusted. Then I reflect on what I currently feel as the result of this imagining: betrayed, downtrodden, and hopeless. Once I identify my thoughts, desires and feelings, I then imagine that that is how Northrup feels as well. In this sense, secondary imagining helps us to identify and empathize with fictional characters. I then remove myself from the simulation, so to speak, and attribute these same feelings to Northrup.

3.2. Complications with ST

ST provides us with a comprehensive psychological foundation for our engagements with fictions if Currie (and others) are right. However, ST also faces several problems as a theory of our social cognition of fictional characters—besides the fact that it appeals to imaginary mental states, which I criticized in the previous chapters. In this section, I will discuss two problems facing ST in general, and two facing ST as a psychological foundation of *fiction*.

First, perhaps secondary imagining—or e-imagining—can help us to discover how a character like Northrup thinks and feels. But in order to achieve this understanding we must first be able to recognize and identify *our own* mental states. As we have seen, high level perspective taking seems to require that we consciously recognize and identify our own mental states through their qualitative feel. This means that we should be aware of our simulation as we do it. Two questions arise here: a) is simulating a part of our conscious experiences with fictions? and b) is it a part of our experience of mindreading others more generally?

I suggested in chapter 1 that we do not consciously engage in a game of make-believe while we watch a film or real a novel. The same goes for simulation; I doubt that most audiences consciously enter into a simulation or go through the simulative process in order to understand how a character feels or what she believe. Of course, the simulation theorist could reply that simulation occurs *unconsciously*. Note that Currie and Walton both take imaginative simulation to be the foundation of our emotional engagements with fictional characters in general. Certainly the simulation should be conscious some, or even *most* of the time. Moreover, how can we square unconscious simulation with the idea that we understand the mental states of others based on our own state's conscious, qualitative character?

Goldman argues that there are some mental states that are necessarily conscious, and it is in virtue of their qualitative character that we can introspect these states (see 2006, §9.6). We can

challenge both of these claims. First, there is significant philosophical and empirical support for the idea that mental states—including qualitative ones, like perceptions and emotions—can occur without conscious awareness (LeDoux 1996, Rosenthal 2005 & 2008, Prinz 2004a, etc.). Goldman argues that our social cognitive capacities generally take place without conscious awareness. But it is hard to see how this is possible for his theory of e-imagining, since e-imagining requires deliberate perspective-taking that we use to recognize our own mental states. The states that we introspect must be available, in some way, to consciousness. How can we square this last point with Goldman's assumption that e-imagining can occur *unconsciously*? It seems that Goldman must either give up the idea that the possession and recognition of our e-imagined states necessarily have a qualitative feel, or that e-imagining can occur unconsciously.

Second, people tend to be quite poor at introspecting their mental states, even for seemingly intense states like emotions (see Carruthers 2010, Mandelbaum 2013, Rosenthal 2008, & Schwitgebel 2007, among others). We often do not know what we currently feel or what our occurrent beliefs are. Goldman seems to be perfectly willing to accept that our introspection is fallible. We may not correctly recognize our own mental state or we may misrepresent the target's. The challenge is to show that our introspective capacities are reliable enough to account for the ease and regularity with which we navigate our social landscape. Some philosophers, though, are skeptical that they are (Carruthers 2010; Schwitzgebel 2007).

So there are problems involved with our capacity to simulate others. Gordon's radical simulation might serve us better; recall that his theory does not require conscious introspection in order for us to understand a target's mental states. This is true; I have not provided knock-down arguments against ST here. Rather, I have considered one aspect of ST that contradicts our actual social experiences: we do not always consciously consider other's mental states by introspecting

our own. My other, more general critique of ST concerns its appeal to imaginative mental states: ST faces the same functional and inferential challenges that "offline" theories of make-believe, imagination, and pretense face.

I also think that there are several problems facing ST as a general psychological basis for our engagements with fictions. Even assuming that ST doesn't face problems concerning conscious experience, it still may be questionable that simulation is required for understanding fictional characters. Recall the connection between simulation and empathy: both require that we put ourselves in another's position, taking on their beliefs, desires, and emotions as closely as possible. Empathy also involves a kind of emotional state, in which we "feel with" another on the basis of imaginatively taking on their perspective. I will have similar emotions to the target if I am successful at imagining her mental perspective—and my emotional state will be caused by hers (Gaut 1999, Hoffman 2008). For instance, I may imagine what it is like to lose a close relative after my best friend's mother passes away. I imagine that I have lost my mother and feel sorrow as a result. My sorrow was caused by witnesses my friends and by empathizing (simulating) her perspective.

Indeed, when people describe our capacity to empathize with fictional characters, they often mean that we undergo some kind of simulative process. Does empathy adequately capture the ways in which we understand fictional characters? If so, then ST might still be a good theory for fictions, if not for social cognition in general.

Noël Carroll argues that empathy is generally both unnecessary and insufficient for explaining our emotional engagements with fictional characters (Carroll 2008). First, our emotional state is generally not type-identical to a character's, even if our emotions are about the same thing (they take the same object). Consider my example from *12 Years a Slave*. I had very

strong emotional responses to the funeral scene. But were my feelings type-identical to Northrup's? Maybe some of them were—both Northrup and I felt sorrow and despair concerning his unfortunate situation. But I mostly felt compassion towards a character that I greatly admire, whereas he felts betrayal, confusion, and hopelessness. I do not feel any of those things; I am not the one betrayed, even if I do feel anger towards Treach. I do not feel hopeless or confused about Northrup's position. Carroll calls this an "asymmetry of affect" (*ibid*, 168); there is a significant difference between the character and audience's perspective, situation, and knowledge. This means that our emotional responses will often be radically different.

Furthermore, even if our emotions *are* type-identical, that does not mean that the character's emotion causes the audiences.' While watching *Star Trek*, both Kirk and I fear the Romulan ship. Our emotions are type identical and have the same object: we feel fear about the approaching Romulans. But that is not to say that Kirk's fear caused mine. My fear was caused by the prospect of a Romulan attack.

Carroll introduces the notion of *criterial prefocusing* in place of ST. Popular film fictions are created to be understood by wide audiences. In order to get such diverse groups of people to understand and respond to fictions, filmmakers prefocus the narrative—they provide us with certain information about the characters, show them in a certain light (literally and figuratively), and deploy features of the medium to guide our thoughts and responses. In fact, virtually every controllable factor of a fiction can be utilized to guide an audience's emotional responses. If we accept the notion of prefocusing, then, Carroll suggests, we generally do not need to introduce the notion of simulation. Our emotional responses and allegiances will already be present without it. According to Carroll, what we have are "vectorially converging emotive states" with a character (*ibid*, 171). Our emotions are often similarly valenced to the characters, but not type-identical to

them. That is, they will both be qualitatively positive or negative. Both Kirk and I experience negative emotions, as do Northrup and I, but not always the exact same type of negative emotions.

I contend that the same basic arguments apply to other mental states besides emotions. We do not generally need to simulate a fictional characters to understand their beliefs, desires, or intentions. Narratives are constructed in such a way that much of the guess work is already done for us. Characters explicitly tell us how they feel or how they intend to act. We are given plenty of information about a character's personality and context so that their beliefs about a situation are often the same as ours—except in cases of dramatic irony, for instance, or when information about a character or situation is deliberately withheld from the audience for some kind of narrative effect. We *may* simulate characters in some cases, but it is generally not necessary to do so.

Carroll rejects ST as the primary type of emotional attachment that we form with fictional characters (2008 & 2011). I think that this rejection is warranted, at least in terms of how we take on the perspective of fictional characters. What about the low level feeling with, such as mirroring, mimicry, etc.? Can ST explain them?

In fact, Shaun Nichols (Nichols et al 1996 & Nichols 2004a) denies that simulation is involved in either low *or* high forms of mindreading. Empathy need not be based on simulative processes, so Goldman and other simulation theorists' equation of the two concepts is misleading. Nichols argues that some empathy is caused by information that the subject already possesses about a particular target. He calls this "information-based empathy." Consider direct association. This holds that we feel with a target when we perceive or hear of her emotion. This triggers an association with a stored representation of one of our own past experiences. Recalling this experience triggers a similar type of emotional response in us. On this view our own empathetic response is at best only *indirectly* caused by the target and may not require any kind of simulation.

Furthermore, Nichols (2004a) is skeptical that mirroring is a kind of simulation (see also Carroll 2008). Mirroring and mimicry are best understood as basic, roughly automatic responses to certain triggers in one's environment, not as a simulation of that person's emotional state. Nichols argues that the latter interpretation assumes a significant amount of folk psychological knowledge that must be internalized before simulation occurs. But, again, Nichols cites significant evidence that mirroring and mimicry occur at very young ages, including in newborns (*ibid*, 64; Decety & Meltzhoff 2011).

Simulation theorists could respond that humans and some other mammals are neurologically hardwired to simulate in this way, and so folk psychological learning is not required for very basic kinds of emotional mindreading. Still, I agree with Nichols that it seems implausible that simulation is at play in mirroring, mimicry, and direct association. We could construe simulation as the representation of another's mental state that then automatically triggers a similar mental state in us. But why think that this is *simulation* as opposed to a brute, associative neurological process? It seems that in order to construe low level mindreading as simulation we must water down the concept to make it significantly less interesting and informative.

I think that we have enough reason to reject ST as the primary explanation of how we attribute mental states to fictional entities. I will now turn to two other prominent theories of social cognition to see how they fare: direct perception theory (DPT) and theory-theory (TT).

4. DPT: Social cognition without mindreading

I will begin by briefly explaining DPT. This theory seems to capture the immediacy of our social

cognitive capacities about fictional characters as well as the theoretical basis of mirroring, mimicry, and association. However, as we will see, this view also faces several problems as a general account of social cognition.

Perhaps we have gone about trying to understand our capacity to understand fictional characters in the wrong way. So far, we have been exploring theories of mindreading: how we think about and attribute mental states to others. But what if we can just *see* a target's mental states?

This is the claim put forth by *direct perception theorists*. Shaun Gallagher (Gallagher & Varga 2013, Gallagher *forthcoming*), a leading proponent of DPT, notes that the current trend in the social cognition literature frames the debate as a problem of accessing another's mind. He calls this the "principle of imperceptibility": other people's mental states are hidden away from outside observers and so are perceptually inaccessible. Attributing mental states and predicting their behavior thus requires the capacity to mindread. The perception of facial expressions and bodily movements may serve as a starting point for mindreading; it is the basic information that gets taken up in inferential or simulation processes. According to ST (and TT, see below), we can only perceive "mere bodily activity, patterns of mechanical movements that warrant the inference or suggest the correct simulation to the other's intentional states" (Gallagher & Varga 2013, 186). Direct perception theorists deny this, arguing instead that perception may go beyond the recognition of mere bodily movement to the meaning or intention behind those expressions and movements. Gallagher & Varga state: "On this account, in our everyday interactions with others, we are able to directly perceive their intentions and emotions; perception can grasp more than just surface behavior—or to put it precisely, it can grasp meaning—the intentional in intentional

behavior and the emotion in emotional expression" (*ibid* 186, 187).[13]

Theory-theorists like Carruthers and simulation theorists like Goldman deny that we can directly perceive another's emotions and goals (see Carruthers 2013 & Lavelle 2012). Our own phenomenological *experience* may be that of direct perception—that is, it *feels* like we can directly perceive the mental states of other people without mindreading capacities. In actuality, though, a great deal of unconscious, rapid simulation or theorizing is required before we can attribute mental states to others.

It may be quite difficult to determine which of these views is correct, at least if we use our own conscious experiences as data points. It does *seem* quite easy for us to glean the emotional state of another person through directly perceiving them, as in Eckman's famous pictures of the six basic emotions (Eckman 1999; see also Plantinga 1999). A curl of the lip, broad smile, or furrowed brow can all indicate a particular emotional state. DPT states that the perception of these expressions is the perception of the mental state itself. This requires that emotions are at least in part constituted by bodily reactions, including (but not limited to) facial expressions. Contrast this position with the weaker claim that we directly perceive the expression of another's emotion, and then unconsciously infer that she is in a particular emotional state. However, if emotions are defined in part by their perceptible bodily reactions, then it makes sense that we can directly perceive another's emotional state. If not, then some kind of inference may be required in order for us to move from a perception of a facial expression to seeing a person's state.

[13] Gallagher and Varga are further committed to the *embodied cognition*, the view that cognitive processes extend beyond the central nervous system to movements and responses from the whole body, and maybe even beyond the physical body to involve one's physical and social environment (see also Noë & O'Regan 2001, Noë 2004; see Block 2005 and de Vignemont 2011 for critique of the embodied thesis). The authors note that one can buy into their DPT without this further commitment.

A further worry, as Gallagher and Varga admit, is that sometimes perception leaves both the type and content of mental states underdetermined. But, they contend, that does not mean that we must appeal to inferential or simulative processes to explain them. Social context and knowledge of a target will also add specification. Unfortunately, the authors do not discuss just *how* social context plays this role in mental state attribution. The challenge would be do show that background information and social context play role in mental state attribution without also requiring mindreading.

There are two further challenges that face DPT, both as a theory of real-life and fictional social cognition. First, what, exactly, does it mean to directly perceive a mental state? Second, can we really eliminate mindreading altogether?

To begin, directly perceiving a mental state seems mysterious. Suppose that we directly perceive a facial expression, such as a smile. On some views of perceptual content (as we saw in chapters 2 and 3), even recognizing a smile requires the inference or judgment that the object we see is of a particular kind. Other theories of perception, such as Susanna Siegel's, hold that we can perceive emotion-kind properties. This means that I can literally perceive happiness in a smile.

Perceiving emotion-kind properties may work for some straightforward cases. But what about opaque situations, or even relatively straightforward cases of social cognition that do not necessarily feature obvious emotional expressions? These situations may require even further inference-drawing and judgments about how another person feels. Indeed, the theory-theorist would counter that "direct perception" actually involves another kind of inference-drawing, this time from a heuristic to a behavior prediction.

Furthermore, Siegel's theory of high level perceptual content is controversial; perception does require inference drawing of some kind on many views. If that's so, then direct perception of

emotional states is not so mysterious: it's just inference drawing from stored information about emotional expressions to a particular perceptual case. However, on this view, the perception isn't so *direct*.

This leads to the second worry: we cannot eliminate mindreading altogether. Gallagher and Varga admit that there will often be situations that do not fit a particular frame and we cannot place a person into a particular role. DPT extends only to emotional expressions and bodily movements that correspond to emotional states and action goals. It does not tell us about mental attitudes such as beliefs, judgments, and desires. These cannot reliably be directly perceived. If that's true, then we may need some kind of high level mindreading capacity to help up in opaque or novel situations. We sometimes need to take the perspective of others in order to understand their mental states. It seems unlikely that we can learn all about Northrup's feelings and beliefs, for example, just by perceiving his face.

At best, DPT can explain *some* cases of social cognition about certain *types* of mental states. It cannot help us in opaque social situations, or sometimes even about the content of mental states in simple ones. Furthermore, the nature of the "direct perception" is mysterious and, quite possibly, inferential. However, DPT does seem to capture the sense of immediacy and automaticity of both our feeling with fictional characters and understanding their emotional states.

5. Theory-theory

The concerns I raised in the previous section about DPT may suggest that we shouldn't eliminate mindreading altogether if we want to know how to understand another's mental states. In this section, I will examine *theory-theory* (TT). TT is often seen as the main competitor to ST.

However, aestheticians rarely appeal to TT as a psychological basis of our engagements with fiction. As stated previously, I suspect that this is because TT does not appeal to our power of imagination, while ST does.

Theory theorists argue that our mindreading capacities—our ability to attribute mental states to others and predict their behavior—stem from our deployment of a folk psychological theory of mind. Some argue that this theory is learned and developed early in life; others argue that it is innate, stored in specific modules dedicated to mindreading. In either case, our theory of mind includes a variety of general social platitudes concerning our folk psychological concepts. These include states like beliefs, desires, intentions, emotions, and perceptions and the typical consequences and basic inferences that arise from their combination. We draw upon this theoretical knowledge and make inferences from particular tenets to a conclusion about what an individual thinks, feels, or intends to do.[14]

TT was born out of work by philosophers such as Wilfred Sellars (1956) and David Lewis (1972), both of whom denied that knowledge of our own mind is the result of direct introspection. Rather, it is a matter of understanding folk psychological knowledge concerning the functional roles of particular states. Sellars intended for this view to combat a Cartesian "Myth of the Given" by which we have direct, infallible access to our own mental states, including both qualitative states and mental attitudes. Similarly, David Lewis argued that folk theory of mind is implicit in our everyday talk about mental states. We regularly discuss and consider platitudes regarding the functional and inferential roles of our mental states. This requires that we have a basic understanding of what kinds of actions and further mental attitudes may result from a particular

[14] See Carruthers 1996 (ed) for a comprehensive overview of different theories of mind and Carruthers 2000 for a discussion of the modularity thesis. See also the Internet Encyclopedia of Philosophy "Theory of Mind" article for a less biased approach.

belief, desire, intention, or pairing of mental attitudes (e.g. belief + desire). We encountered this basic functionalist idea in chapter 1. We have a basic understanding of the sorts of behaviors that will result from our occurrent beliefs and desires.

Contemporary theory theorists include Alison Gopnik (1993), Jerry Fodor (1983), Peter Carruthers (1996a), Alan Leslie (1996 & 2000), and Simon Baron-Cohen (1995). Each of these theorists argues that our social cognitive abilities arise from a capacity for utilizing folk psychological knowledge and drawing inferences about our social environment based on that knowledge.

The two dominant versions of TT are the "child-scientist" theory held by Gopnik and the modular theory favored by Carruthers, Baron-Cohen, and Fodor. The difference between these two versions of TT concerns how we acquire our folk psychological theory of mind. Gopnik argues that this knowledge is internally represented by normally functioning persons and has much the same structure as a scientific theory, including theories, postulates, and observations. Our theory of mind is used in much the same was as scientific theories are, as bases for our own observations and methods to help guide our inferences. The theory starts out early in life as quite limited but develops and becomes more sophisticated at regular periods throughout a child's cognitive development—as seen, for instance, in the child's ability to pass theory of mind and false-belief tasks around age four, but not previous to that age (Gopnik 1993; Gopnik & Schulz 2004; see Wimmer & Perner 1983 and Harris 2000 for overviews of the task).

Carruthers and other modularity theorists find the child-scientist view untenable. He points out that it is remarkable that almost every child develops roughly the exact same theory of mind at the exact same stages of development. The child-scientist theory lacks the precision to explain these regularities in development—on that view, why wouldn't individual children develop quite

different theories based on their unique environment and personal idiosyncrasies? The best way to explain developmental regularities is by positing an innate theory of mind module, a psychological mechanism that represents certain specific information and may be separate, or encapsulated, from other modules, so that each module specifically pertains to a particular task (Fodor 1983). The theory of mind module gradually becomes more functionally activated throughout development, similar to a Chompskian notion of a language module that allows us to acquire a first language at regular intervals throughout our development.

Like ST, TT also developed from research on the childhood development of social cognition, and particularly from research on children with ASD. The cognitive, imaginative, and social deficiencies that plague children with ASD may be the result of a deficient theory of mind, whether learned or innate (see Gopnik & Schulz 2004 for the former position, Baron-Cohen 1995 & Carruthers 1996b for the latter; see Currie 1996 and Goldman 2006 for a denial of theory of mind in ASD). These children often have great difficulty in navigating their social environment because they lack the typical folk psychological knowledge that is required for maneuvering their social environment. Contemporary theory theorists include Alison Gopnik (1993), Jerry Fodor (1983), Peter Carruthers (1996a), Alan Leslie (1996 & 2000), and Simon Baron-Cohen (1995), each of whom holds that our social cognitive abilities arise from a capacity for utilizing folk psychological knowledge and drawing inferences about our social environment based on that knowledge.

Maybe we do possess a great deal of folk psychological knowledge. But it certainly doesn't *seem* like it during our actual, first-person experience of our social interactions. Surely we generally don't deliberately or consciously draw inferences about another's mental states and behavior. Again, mindreading often seems to take place automatically, without deliberation,

inference, or any kind of conscious theorizing. Carruthers believe that a modular TT can accommodate this first person experience. From an early age, we internalize a great deal of knowledge about the mental lives of others, supplementing this knowledge to our innate theory. We practice implementing this theory on a regular basis, eventually becoming so good at mental state attribution that it seems automatic and can occur very quickly. Furthermore, much of the mental processing that occurs during this process takes place a subpersonally and unconsciously. We may only be consciously aware of the end result. Consider, for example, how I come to be consciously aware of a tree outside my window. I have no knowledge of, or access to, the variety of processes and mechanisms involved in my perception. I am only consciously aware of the *product* of all that processing. The same may be true of our social cognitive capacities; the knowledge of our own and other minds may *feel* automatic, but it is always the result of some kind of theoretical inference (Carruthers 1996a, 27).

So TT is able to sidestep the worry concerning consciousness that plagues ST. However, TT is not typically used as a tool to explain our engagements with fictions. This is surprising, considering the vast literature on fiction and simulation.

Mental theorizing about fictional entities would arguably be roughly the same as how we theorize about real people. There may be some differences in terms of the intentional *content* of our theorizing between real-life and fictions. We never have the experience of traveling through space and fighting Sith Lords, for example. There may also be some limitations in terms of what we have the folk knowledge to theorize about; e.g., some situation that is wholly unfamiliar to us in the real-world. But this is no different than if we encountered an unusual or unfamiliar person or circumstance in our daily lives. Finally, mental theorizing about fictional entities may require a reader or viewer to understand the conventions of a particular fictional media. I will return to these

points momentarily.

So far, TT sounds quite compatible with taking the fictional stance. This makes it worth exploring as a psychological foundation for the SAV. The key to understanding how we mindread fictional characters requires us to understand how we do so in real-life cases.

Let's begin with an example. Suppose that you are having dinner with some friends. Everyone at the table is laughing merrily, eating and drinking with gusto. Everyone, that is, except for your best friend. She sits at one end of the table, barely eating and not joining in with the others' frivolity. You want to understand how your friend currently feels (remember that this can take place consciously and deliberately, but generally will not). Your perception of your friend—her slouched shoulders, crossed arms, tense facial expression, quietness, untouched plate of food and glass of wine, etc.—all serve as data points that you use in your inferences to draw a conclusion about her feelings and beliefs. Perhaps your folk psychological knowledge tells you that, generally speaking, *that* kind of facial expression and *those* kinds of behaviors often correspond to an anxious emotional/mood state. Deploying your theoretical knowledge about emotions, you may put together all that you perceive about your friend, as well as some information about social context and your friend in general, and conclude that she is anxious about something.

We can apply this basic TT framework to our engagements with fictions. We do not project or imagine ourselves in Northrup's situation in order to understand what he is feeling. All we need to do—after taking the fictional stance—is call upon our folk psychological knowledge of emotional expressions. This squares with my discussion of Northrup in §2.1. On this view, I watch Northrup's face at the funeral and draw inferences from my knowledge of facial expressions and Northrup's relevant beliefs and desires to conclude that he feels bewildered, angry, sorrowful. I do not need to imagine myself as Northrup, a slave in the Deep South who was recently betrayed by

a man named Treach, in order to gain access to what he is thinking and how he feels. This also implies that if I feel *anything* towards Northrup, it will not necessarily be the same emotion-type as he feels, but rather whatever emotion that one would experience upon learning about Solomon's situation, which may be similar to Solomon's but not type-identical.

There's another positive feature of applying TT to our social cognition of fictional entities: TT can explain how our knowledge that an object is fictional affects our responses to it. Understanding fictional entities requires that we have some knowledge of fictional and medium-specific conventions. Recall that when we take the fictional stance that we do not forget that the object of our engagement is non-actual. For example, we know that Northrup is a fictional character and does not exist in our world (even if he is based on an actual person). This knowledge carries with it certain information. We know that individual time-slices of Northrup are supposed to correspond to one unifying fictional person. We know, if we are watching a film, that we cannot reach out and touch Northrup. We also know that Northrup's story will probably follow certain conventions. We know things about his past, although not as much as we would if we were to be as emotionally close to an actual human embodiment of Northrup. Basically, knowledge of fictions may be developed and employed as a special kind of theoretical knowledge, knowledge that is fairly easily acquired for almost every human. If we consider our engagements with fictional characters a kind of social cognition, then it may be that our knowledge of fictional characters to be a highly specialized kind of mental theorizing.

According to TT, mirroring, mimicry, and direct association may not be a matter of simulating another's mental state, which implies that we ourselves must copy the observed state of another. Instead, we take our perception of another's facial expression or movements as data points to be explained. A theory-theorist would argue that, once we deploy our theoretical

knowledge about facial expressions, movements, and intentions, our own motor and emotional systems are triggered, leading to the kind of effects that mirroring/mimicry researchers have reported. Direct association implies that we would draw comparisons between our own experiences of a particular type of emotional situation and what we felt in that instance, and then infer that the other person must feel the same way. Or, as Carruthers (1996a & 2010) points out, even if these low level mindreading does involve simulation, theoretical knowledge and inference-drawing may still be required for us to get from the perception of another, to a representation of her emotion, to our own simulation of that emotion, and back to mental state attribution. Granted, these inferences and knowledge will be subpersonal and unconscious—they will not feel at all like deliberate inference-drawing. The important point is that, *prima facie* anyway, folk theory and inference-drawing can explain low level feeling with without appealing to simulation.

High level feeling with may pose more of a problem if it is necessarily imaginative. Recall that simulation theorists argue that perspective taking involves imagining ourselves to be in the situation of another person. For a short time, I imagine myself to be my friend, with all of her relevant properties. I then conclude that she feels a certain way based on what *I* imaginatively feel. In contrast, Carruthers and other theory-theorists argue that perspective taking draws upon a great deal of knowledge about minds and mental processes. The ability to recognize the mental states of another based merely on facial expressions and body language seems to require us to have a great deal of knowledge about bodily expressions and behaviors. We also need to draw inferences from what we know about emotions and our simulated imagining to a conclusion about the mental state of another.

Perhaps both high and low level feeling with are primarily associative processes, and inferring another's mental states on the basis of those. Inferring the emotional state of another may

be an after effect of recalling situations that are meaningful to us. Do basic mindreading capacities, like mimicry and mirroring, require inference drawing? The theory-theorist argues that they do (Carruthers 2013; Lavelle 2012). On their view, even these basic perceptual and affective processes are inferential—for example, we must draw an inference from the perception of a face and our stored representations of those kinds of faces in order to conclude that the current perception is of a certain kind.

One potential issue with this explanation is that TT is supposed to appeal to *folk psychological* knowledge—i.e. Lewis's mental platitudes. Like other critics of TT, I find it implausible that our basic knowledge of mental processing includes information concerning perceptual and affective processing (see, for example, Currie 1996 & Heal 1996). If that's true, then TT faces a serious challenge regarding low level mindreading.

There's a further worry about TT in general that is worth noting. The examples of mental theorizing that I have used up to this point have been fairly general. *In general*, we can learn of a target's mental state type by theorizing from our (innate or learned) folk psychological knowledge about behaviors expressions, body language, social context, etc. We then draw a pretty good inference concerning the generic mental state the observed person is in. So TT may give us a fair idea about general mental state types. But can theorizing from general knowledge and platitudes tell us the *content* of a particular state? Some critics are skeptical that it can. Jane Heal (1996; see also Perner 1996) makes this point. She notes that TT faces a problem similar to the 'frame problem' in computer science. Very briefly, this implies that we have a limited capacity to store and deploy folk psychological knowledge, but our social interactions require a vast amount of stored information about the mental lives of others and ourselves that we deploy on a regular basis. Our innate or acquired theory would have to be incredibly complex to deal with any particular

situation in which we find ourselves.

Part of the problem is that the folk psychological theory has to be general enough to apply to a great deal of situations. But then we face another problem—the theory is not specific enough to be able to help us understand the content of another's mental state at a particular moment. If we really want to understand Northrup's mental states, for example, we not only need to be able to explain what type of emotion he feels (for example), and *why* he feels that way. Perhaps a folk psychological theory could explain the content of a mental state, but then the theory would be far too complex to be realistically mentally stored. According to theorists like Heal, Perner, and Nichols (Nichols et al 1996), the frame problem suggests that TT must either be discarded or supplemented by another, simpler mechanism, such as a simulative one.

6. A Modified TT

I am sympathetic to the critiques raised concerning both TT and DPT. However, I also think that TT is the best kind of theory for explaining how we understand the mental lives of fictional characters. Moreover, DPT captures the prima facie immediacy of our ability to perceive the mental states of others.

In this section, I will spell out my own account of mindreading fictional entities: a modified version of TT that, I hope, makes up for its downfalls. I achieve this by supplementing the standard TT model with a theory of *social referencing*, another way in which we attribute mental states to others that does not require high level mindreading. However, it still requires inference drawing, as we will see.

6.1. Social referencing

José Bermudez' "social referencing" model contends, like Gallagher's DPT, that social cognition often does not require mindreading *per se*; rather, it is often enough that we understand certain social scripts, roles, and schemas and make use of basic cognitive heuristics to understand a target's mental states and to predict her behavior (Bermudez 2005).

The idea that cognitive processes involve heuristics has become pervasive in cognitive science. A traditional view of human cognition is that we are "rational actors"; we choose what options to pursue by assigning and assessing the probability and results of each possible outcome, determining the utility of each to the best of our ability, and then combining these assessments. The option we choose—either for our own actions, judgments, and beliefs, or for understanding another's—is the one that maximizes probability and utility (Gilovitch and Griffin 2002). But there is a decent amount of recent evidence that humans do not always think like this (see Nisbett & Wilson 1977, also Damasio 1994 & Sloman 2002). We take shortcuts that minimize the amount of cognitive strain and energy required to reach a decision, judgment, or solution to a problem. We have "bounded rationality"; there are processing limitations to the human mind (Simon 1955). People may reason and choose rationally within the constraints imposed by their limited search and computational capacities. We are also "cognitive misers." We utilize as little cognitive power as possible that is needed to efficiently reach some conclusion or result in some action (Eiske and Taylor 1991). We are not cognitively lazy. Rather, we have developed problem solving shortcuts that generally allow us to quickly and effectively maneuver our social environments, freeing us to pursue other activities.

One method of achieving this comes from the implementation of *cognitive heuristics*. Very basically, a cognitive heuristic is a problem solving capacity that allows us to make a judgment or reach some conclusion by utilizing simple devices in lieu of more costly reasoning processes. For example, Walter Sinnott-Armstrong and his colleagues define heuristics as unconscious attribute substitutions: we determine that *X* has a target attribute by unconsciously using information about another, different attribute that is easier to detect (Sinnott-Armstrong et al 2010). For example, by using the *availability* and *representativeness heuristics*, one forms a belief about a relatively difficult attribute of a situation or object based on a more accessible one. In one study, participants were asked many seven-letter words whose sixth letter is 'n' (_____n_) occur in the first 10 pages of Tolstoy's *War and Peace*. They were also asked how many seven-letter words ending in "ing" (____ing) occur in the first ten pages of *War and Peace*. Interestingly, the average answer to the first question was significantly lower than the answer for the second. But, as Sinnott-Armstrong and his colleagues note, the correct answer to the first question cannot possibly be lower than the answer to the second since every seven-letter word ending in "ing" is also a seven-letter word with 'n' as the sixth letter! Participants tend to make this mistake by using the availability heuristic—their guesses are based on how easy it is for them to come up with examples, and some examples are more readily available to us. It is significantly easier to come up with examples of the seven-letter "ing" words; less so for seven letter words with 'n' in the sixth spot (*ibid,* 249).

The *recognition heuristic* is similar to the mere exposure effect, according to which we choose and tend to prefer things that we know over those with which we are unfamiliar (Zajonc 1968). An important heuristic for our purposes is the *affect heuristic*. Thinking about an action or person, whatever it is, may make us feel positively or negatively. We take these feelings in turn to constitute judgments that the action or person is morally wrong (Sinnott-Armstrong et al 2010;

Slovic et al 2002). Our moral beliefs are based on these emotional responses instead of deliberate reasoning processes. I will return to a discussion of the affect heuristic in chapter 6 when we discuss our moral judgments of fictional entities.

Each of these heuristics may provide us with quick, efficient means of judging a situation, whether we are in a social context (as with the affect heuristic) or trying to solve some problem (as in the availability and representativeness heuristics). There is a sense in which simple heuristics "make us smart"; they allow us to come up with judgments that we wouldn't normally be capable of making, or at least not as quickly or easily (Gigerenzer et al 2002). On the other hand, cognitive heuristics are fallible and, because they are not always under our deliberate control, they can be unduly influenced by biases or implemented without sufficient information (see Kahneman & Tversky 2002). Still, many researchers agree that cognitive heuristics are accurate enough in typical environments; problems arise in difficult or atypical ones.

Returning to social cognition, authors like Bermudez hold that we can utilize cognitive heuristics to help us understand our social environments. We do not always need to go through cognitively expensive high level mindreading processes to understand the mental states and actions of others; rather, we can make use of simple generalizations and rules that will key us into what another is most likely thinking. Bermudez highlights two ways in which we can utilize heuristics for social cognition. We have already encountered the first in terms of Gallagher and Varga's DPT: emotional perception. Our perception of the emotional state of another person depends on cues outside of conscious awareness. We recognize these cues, such as facial expression and body language, without consciously or deliberately identifying them. This perception acts as a kind of heuristic by which we attribute a more complex mental state to another based on something that is relatively easy for us to recognize (Bermudez 2005, 199).

Second, we can base more complex social interactions on simple social rules, scripts, and frames. Bermudez asks the reader to imagine that you are sitting in a restaurant, about to order lunch. A server walks towards you. We can ask ourselves what the server is thinking at this moment and why he acts as he does (walking to the table). Do we need to go through a mindreading process to answer these questions? Bermudez suggests that we don't. Instead, we can put into practice information that we already possess concerning our current social situation and social roles we fill. In this case, you are in the 'diner' role. There are standard norms of behavior for how to act in this role: you sit down, you are quiet, you do not shout across the restaurant to grab the server's attention, etc. There are also typical mental attitudes: intentions, desires, beliefs, etc. The server also fills a role: his job is to seat you, bring you what you need when asked, take your order, etc. Together, this information makes up our restaurant 'frame'—a general template for a particular social situation—and your server follows a general behavioral 'script.' To answer our questions about what he will do and what he is thinking we do not need to engage in mindreading. We just need to recognize and apply the right frame to the situation and roles for the person. In this instance, we would probably be correct in determining that the server is walking towards me because he wishes to take my order.

How would social referencing help us to understand the mental states of fictional characters? Carroll's own view appeals to something like Bermudez's social referencing. Carroll's view of how we understand the mental states of fictional film characters, the *cognitive heuristic model*, blends DPT and social referencing (Carroll 2008). Carroll discusses the heuristic model in light of complications with both ST and the TT. Instead of simulating the mental states of another person in order to understand their mental states (as in ST) or employing a mini scientific theory to do so (TT), Carroll suggests that we glean the mental states of others on the basis of schemas,

encoded scripts, prototypes, contextual cues, exemplars, and other heuristics (*ibid*, 174-175). He notes that:

> When we hear that a relative has secured a long-sought-after job, all things being equal, we suppose that she is happy and we rejoice for her. There is no need for simulation. We have access to a body of prototypes regarding emotional responses in certain contexts as well as recognitional cues, such as facial expressions and postures, which enable us to assess the emotional states of others. These are not theories and they are not applied by subsuming particular situations under nomological generalizations as the theory-theory might have it. Rather, they are prototypes—schemas, scripts, and recognitional cues—employed by *analogy* (analogy rather than subsumption) (*ibid*, 175).

Like Bermudez, Carroll suggests that frames, heuristics, scripts, roles, etc. are enough for us to understand the mental states and predict the behaviors of others. This is especially true in situations like the restaurant, according to which there are clear normative guidelines for how one should behave. Other examples come to mind: our behaviors and feelings at sporting events, in museums, on the subway, in a doctor's office, at the family dinner table, etc. Perhaps Carroll is right in saying that fictions present us with even greater range of scripts and frames. We can generally guess, based on prefocusing techniques as well as the narrative, how a character will act, how she feels, and what she is thinking even without engaging in mindreading. Much of the information has been provided to us by the filmmakers.

However, social referencing also faces similar challenges to DPT. First, it's not entirely clear that the use of scripts, frames, etc. is non-inferential. Second, the scripts and frames we do possess may not help us in ambiguous, opaque, novel, or complex social situations. Any social behavioral that is run "off-script" will not be understood. It seems, then, that we cannot completely do away with mindreading.

Consider our attempts to understand Northrup. This scene is opaque; we need to engage in

some kind of mindreading in order to fully understand how Northrup feels. Indeed, our social frames may not be specific enough to apply in each particular scene that we encounter. We may not possess the requisite knowledge of which scene to apply in a specific case. For *12 Years a Slave*, we would need to have a fairly general script about the social interactions between slaves and masters, and slave life in general. Even if we do possess these frames, it is not clear that even this script would apply in Northrup's case. We would need to supplement our heuristic abilities with mindreading in order to fully understand how Northrup feels.

6.2. TT with social referencing

Neither TT nor social referencing can explain the full range of our social cognitive capcities on its own. I propose combining the two theories to form a modified TT that makes up for the challenges each face on their own. Recall the explananda for a theory of social cognition of fictional entities. We often find ourselves in social situations where we consciously desire to know more about what another thinks, feels, or intends to do. Our social cognitive capacities seem to occur automatically, without telling them too. It's as if we are constantly on the lookout for social cues concerning what other people are thinking and feeling and our social cognitive capacities are deployed without forethought. Low level mindreading techniques may help explain how this is possible; processes like mirroring and association take place without deliberation or thought. In other occasions, we may need to consciously and deliberately engage in high level mindreading, like perspective taking. This would help us in ambiguous and novel social situations.

I think that we can account for these explananda with a modified TT. On the one hand, we think of ourselves as fully rational, deliberate thinkers that consciously work through steps of a

problem, weigh potential outcomes, and draw conclusions. TT has a ready explanation for this. We make associations between our current social climate, including the mindreading target, and our folk psychological theories about how people typically act in such situations. We draw a conclusion about what a target thinks or feels based on our beliefs, contextual information, and knowledge about the type of situation at hand. This includes a target's behaviors and bodily expressions.

On the other hand, it seems quite clear that many thinkers do not typically solve problems this way, or at least not consciously. Quite a bit of our problem solving occurs unconsciously and automatically, utilizing quick and efficient cognitive tricks and tools that maximize efficiency and minimize cognitive expenditure. TT can help us here as well. As I suggested in 5, even low level mindreading, such as mirroring and mimicry, may require some kind of inference drawing from tacit theories about social interactions and mental states. Even the ability to recognize facial expressions and conclude that a target feels happy, sad, or anxious requires inference from stored information to a particular case. On this view, "direct perception" isn't that direct. Our ability to perceive emotional and/or intentional states may actually involve some kind of inferential judgment, even if it doesn't consciously feel like it.

The problem was that TT seems overly complex. It requires us to hold a great deal of tacit information about social interactions. Heal compared this to the frame problem in computer science; the human mind cannot store as much data as TT seems to require in order for it to be specific enough to account for complex social situations and to reveal detailed information about the type and content of a target's mental states.

But what if TT *doesn't* require us to have detailed information about particular social situations? What if there is a cognitively "cheaper" version of it? This is where social referencing

enters the picture. At least some of our tacit knowledge about social situations comes in the form of mental heuristics like scripts, frames, and social cues. When we find ourselves in commonplace or familiar social situations, we draw upon those scripts and frames in order to understand how the people around us think and feel. We draw inferences from our tacit scripts and frames to what we perceive and believe about our current circumstances. We then infer what a target will do, how she feels, or what she believes on the basis of this knowledge.

Understanding and predicting behaviors will often require us to draw upon perceptual capacities, background information we have about the particular person, *and* mindreading. It may be that in stereotypical situations we can call upon social referencing cues, frames, and scripts to explain another's behavior. Low level mindreading and social referencing help us here.

In other cases, particularly opaque or challenging social interactions, our seemingly immediate mental state attributions will have to be supplemented by more deliberate social cognitive abilities, like inferential perspective taking. Sometimes social referencing will give us a good idea of how a target feels. However, we sometimes want to delve deeper into another's perspective or we just can't tell what to think of her. This will require further, more cognitively sophisticated, mindreading based on our tacit folk theoretical knowledge.

Let's reexamine the scene from *12 Years a Slave*, this time in terms of my modified TT. Upon perceiving Northrup's own highly expressive face, I come to recognize that he is in a negative emotional state, such as deep sorrow. This judgment may arise from low level mindreading processes like mirroring, inferred from tacit knowledge about what facial expressions reveal about emotional states. I may also understand that Northrup feels betrayed and believes that his situation is hopeless. I will acquire this knowledge based on the information I have about this character and, again, what his behavior and facial expression reveal about his inner states.

As Northrup's expression changes, so too does my own interpretation of what he feels. I may infer (through either low or high level mindreading) other kinds of emotions in his face, such as anger. At the same time, I also consider what I know about Northrup and what I know about the history of the American slavery, any other information that might be relevant to my understanding of what Northrup feels, and why he feels that way. I may be surprised and perplexed by his sudden look towards the heavens and then need to take on his perspective momentarily to understand why he did this.

I argue that both low and high level mindreading processes inform our knowledge of Northrup's mental states. Both are inferential, some resulting from tacit knowledge of social scripts and frames, some from more general folk psychological knowledge about human behavior and mental states. Ultimately, I argue that this view avoids the problems that plague each of the individual theories that we have previously encountered, and especially ST. We do not need to posit distinct mental states, like imaginative beliefs and emotions, in order to understand how a target feels. This is true for both real-life and fictional objects. All we need is our tacit folk psychological knowledge about mental states and social situations.

This chapter built off the promise of the previous one. By taking the fictional stance, we both recognize fictional entities *as* fictional and as the kind of thing that we would normally attribute mental states to. That is, we see fictional representations of people as people, possessing beliefs, desires, and emotions. At the end of the day, I do not think that there is much of a difference between how we attribute mental states to real-life people and fictional characters. The main difference, of course, is that the content of our mental states towards fictional entities will contain a fictional operator from taking the fictional stance. That is, I recognize that this is a fictional character that I can respond to. I will make similar claims concerning our emotions and moral

judgments about fictional entities; we use roughly the same cognitive capacities and processes when we feel for and morally judge both fictional entities and real people. Differences in reactions will be the result of differences in the content of our mental attitudes.

Chapter 5: Genuine, Rational Fictional Emotions

1. Emotions & fiction

One of the main reasons why we engage with fictions is to experience certain emotions. We enjoy feeling fear while watching a horror flick. We revel in experiencing the joys of a new love along with our favorite characters. We are excited to "visit" interesting and foreign locales. This chapter addresses the nature of these emotions. In chapter 1, I argued that the three main arguments in favor of the DAV—the arguments from function, inferential role, and neuroscience—are unable to establish that we utilize a distinctive mental state in our engagements with fictions. Many contemporary aestheticians argue in favor of the DAV for emotions even if they do not explicitly do so for other mental states. Perhaps there are unique features of emotions that we have not yet covered that would justify this move.

We individuate mental states, including emotions, according to various components. This includes emotional feelings, emotions' impact on cognition and behavior, and emotions' semantic and normative implications. Each of these favors is potential fodder for the DAV. If it can be shown that the feelings, functional role, etc. of a fictional emotion are different than those involved in emotions about actual objects, then it could be that we do not have stereotypical emotions about fictional ones.

The paradox of fiction, one of the puzzles of fiction we encountered in chapter 1, famously embodies the debates concerning the nature of our emotions about fictional objects. Here I show how we can dissolve the paradox by questioning the functionalist assumption inherent in the

typical responses to it. These assumptions also motivate the DAV; without them, the DAV loses much of its theoretical pull.

My goal is also to understand, as best as possible, how and why we experience the emotional responses towards fictions that we do. So my discussion of fictional emotions will go beyond the paradox to a discussion of the nature of emotions. I discuss the paradox of fiction in §2. In §3, I will present a multi-level appraisal theory of emotions that I believe can best explain our emotional responses towards fiction. It is also fully compatible with the SAV and the fictional stance. Finally, in §4 I will discuss the semantic and normative implications of our emotional responses towards fictional objects. I will attempt to thwart one final potential argument in favor of the DAV for emotions: our emotional responses towards fictions are not genuine because genuine emotions have certain semantic and normative implications, and those towards fictions do not. I will argue that we can have genuine, rational emotional responses towards both real-life and fictional objects. I will summarize my conclusions from this chapter in §5.

2. Genuine fictional emotions

2.1. The paradox of fiction

The paradox of fiction offers a succinct summary of the claims involved in the DAV/SAV debate concerning emotions. Cognitive belief-based theories of emotions were in full-sway when the paradox was first introduced (Radford 1975, Walton 1978, Currie 1990). According to these views, an emotion about an object *X* requires that we have some relevant belief *Y* concerning *X*'s relation to our well-being. For example, experiencing fear requires that I believe that there is an

object in my environment that could harm me or someone I care about. We lack the emotion if the relevant belief is absent (Solomon 1993).

The wording of the paradox reveals an adherence to a belief-based theory of emotions. The paradox states:

1. We have genuine emotions about fictions all of the time.
2. We do not believe that fictional characters exist.
3. We can only have genuine emotions about things we believe to exist.

The paradox captures a very natural thought concerning our emotions: if we know that we are engaged with a *fiction*, then we should not have the emotionally relevant belief. No emotion *should* arise. Nevertheless, we have emotional experiences with fictions all the time, whether these are genuine emotion-states or not.

One interpretation of the paradox states that there is something fundamentally *irrational* about our responses towards fictions. Colin Radford (1975) accepts each of the paradox's propositions, notoriously arguing that our emotions towards fiction force the reader into adopting two contradictory beliefs: we both believe and do not believe that the fictional object of our emotion exists.

Alternatively, it could be that our beliefs concerning fictions are different *kinds* of beliefs than those we possess about the real world. This is the motivation behind the DAV: there are significant differences between our emotional responses to fictional and real life objects. The difference is so great that they are actually distinct kinds of mental states. Thus, a proponent of the DAV would solve the paradox by denying the first proposition while maintaining the cognitivist position that emotions are constituted by beliefs.

Few contemporary philosophers opt to eliminate the second proposition. Doing so implies

that a reader or viewer of a fiction would actually believe that the fiction is real while she reads or watches it (Coleridge 1985, Hurka 2001). Theorists who opt to reject the second proposition of the paradox must somehow square the "suspension of disbelief" with the contradictory beliefs and actions we seem to have in response to fictions. This proposal may work if we accept the cognitive illusion thesis, the idea that we are deluded into believing that fictional content is real during our engagement with it. I have already argued against this position in chapter 2.

Many philosophers opt to eliminate the third proposition. There are several ways to do this. First, one can deny that beliefs are a necessary component of emotions, but still maintain a cognitivist position that emotions are comprised of thoughts (Carroll 1990 & Lamarque, 1981) or judgments (Solomon 1993). For example, when we engage with a fiction, we generally have various thoughts about the characters. While watching *The Conjuring*, I may contemplate the nature of the demon that possesses one woman. This thought fills me with terror. Importantly, thoughts do not have the same assertoric requirement that beliefs do. We do not need to believe that the object of our thought actually exists in order to contemplate and respond emotionally to it.

Alternatively, one can deny that *any* higher-order cognition is required for emotions. This is the route taken by non-cognitive perception and feeling-based theories of emotions. According to these views, an emotion does not require that we have a thought, judgment, or belief about an object in our environment. Conscious feelings, bodily changes, or perceptions of those changes, constitute an emotion (Goldie 2000, James 1890, LeDoux 1996, Prinz 2004a, Robinson 2005). The ontological status of the emotion's object is more or less irrelevant to whether or not the emotion itself is a stereotypical state; if the feeling or perception of bodily changes is genuine, then the emotion is as well.

My own theory of emotions is in line with rejecting the third proposition. I argue that

beliefs, judgments, and thoughts are not necessary components of emotions. However, it is not my goal here to solve the paradox by showing that one of the propositions is false. Instead, I argue that the proponent of the SAV should reject the paradox altogether. Philosophers and psychologists typically attempt to dissolve the paradox by appealing to one or another theory of the nature of emotions. They argue that emotions are constituted by beliefs, thoughts, feelings, etc., and show how this eliminates one of the propositions. My approach is slightly different. I think that most philosophers writing on the paradox already assume the DAV assumption that we are only motivated to act by our encounters with real-life objects in our environment. I already developed a response to this claim in the first chapter by reevaluating what we mean by motivation towards fictions.

The same type of argument also applies to emotions. One benefit of this approach is that it applies to *any* theory of emotions—cognitive or non-cognitive—because each holds that emotions have a functional role in virtue of their constitutive cognitive or bodily component. So in order for the DAV to be correct, one would have to show that our emotions towards fictions do not play the right sort of functional role as real-life emotions do, just as beliefs towards fictions purportedly do not.

What is the functional role of emotions? The best candidate would be the actual expressive and behavioral responses that our emotions elicit. Emotions are often associated with certain action tendencies. These include such reflexive and automatic behaviors as a fight or flight response, freezing, running, screaming, or covering one's eyes in fear, clenching one's teeth or striking out at a foe in anger, or smiling, laughing or jumping for joy. We might also add more complex action tendencies or dispositions, which are caused by an emotion but require further beliefs, thoughts, desires, or judgments in order to be manifested. The moral emotions might be included here: we

shun, blame, and punish those who are the objects of our contempt or anger. We comfort, protect, and in other ways care for those who are the objects of our pride, pity, or compassion. If we feel shame or embarrassment, we may withdraw from society or attempt to make amends for our wrongdoings.

Consider an emotional response to an act of injustice. We learn that a seemingly innocent person has been falsely accused and convicted of a crime. We emotionally evaluate the situation and conclude that the court's decision was unjust. Our indignation motivates us in several different ways. First, we may express our anger via our facial expression and bouts of shouting. Second, our indignation plays a role in our further cognitive processing, shaping, influencing, and regulating later thoughts and decisions. Finally, our anger may help motivate us to *act*: perhaps you organize a peaceful protest, sign petitions, or reach out to a local member of congress. Or maybe you simply condemn the court's decision amongst your close friends and family. Your emotional response to the situation motivates you in each case.

However, if the innocent person you care for is actually a *fictional character*, it seems like you cannot be motivated to act in any of these ways. This leads us straight to the heart of the matter for supporters of the DAV. Our emotions towards fictions are not genuine because they do not display the right kind of behavioral output. They do not lead to the same sorts of actions as a genuine emotion would.

I contend that it is possible that we are somehow motivated to act on our emotions towards fictions. If my favorite character in a film dies, I may be strongly motivated to leave the theater, turn off the television, or close the book and go mope in my room for an hour, just as I would if I learned of the death of a popular politician, sports figure, or musician. A complex fictional story may provoke a thoughtful conversation between friends, just as the unjust court decision does. It

is also possible that we are in some ways motivated to act on the basis of our emotions about fictional objects, but other information and circumstances—especially the knowledge that the object of our emotion does not physically exist in our world—blocks or regulates the action. We may be automatically and unconsciously primed to act, but contradictory beliefs and judgments may put a halt to that action before it reaches fruition.

So it is possible that we are sometimes motivated to act on our emotions about fictional objects. This does not require that we suspend our disbelief about the reality of a fiction; rather, we must reconsider what we mean by emotional motivation. We can also side-step the paradox of fiction. Most responses to the paradox buy into the functionalist argument that we will only act in light of things we believe to exist. My view shows how we can diffuse the paradox without that assumption, by eliminating the functionalist motivation that inspired it.

This would be a major win for the SAV. There is nothing about our emotional behaviors that suggests that our emotions towards fictions are not stereotypical states. The proponent of the DAV could continue attempt to show that some *other* feature of emotions makes them distinct from genuine emotions. I will consider four more emotion properties that might be different in our emotions about fictional and real objects: inferential role, emotional feelings, emotional evaluations, and issues concerning emotions' semantics and normativity.

2.2. Emotions' inferential role

The proponent of the DAV can argue that our emotions about fictional objects serve a different inferential role than emotions that are about real things. Perhaps the distinct inferential role is enough to constitute a different type of state.

We must first try to understand the inferential role of emotions. In order for emotions to play a role in cognition they must have some sort of cognitive content. What might that be? If emotions are constituted by propositional attitudes, then we could say that their inferential role arises due to this. For example, if an emotion is constituted by a thought with certain content—a thought that, say, "it really irks me when my sister forgets my birthday"—then the emotion's inferential role would be determined by the role of that thought. The same could be said for cognitive judgments and beliefs. The story would be slightly different for theories of emotion that do not hold that emotions are constituted by propositional attitudes (e.g. feeling theories and perceptual theories). But even these theories tend to accept that emotions have intentional objects; they represent something or someone in the subject's mental or physical environment. An emotion's inferential role stems from this representation. Let us suppose, then, that an emotion plays a certain inferential role in virtue of its *intentionality*, however understood.

An emotion's intentional object may be something out in one's environment, such as a person, animal, or state of affairs that we witness in real-life or while engaged with a fiction. The intentional object may also be another mental state, such as a belief, thought, desire, judgment, or perception of the environment or one's own body. The emotion may in turn play a role in causing or influencing further mental attitudes. For example, suppose that it's your best friend's birthday. Everyone in your friend's family has called to wish her well—everyone except her older sister. The sister's carelessness is both the cause and intentional object of the great annoyance you experience on your friend's behalf. You must have other mental attitudes in order for you to undergo this emotional response: the desire for the sister to call, the desire for your friend to have a nice birthday, the belief that a sister should call her sibling on her birthday, the belief that the sister has not done so, etc. All of these attitudes contribute to your current emotion. Furthermore,

your annoyance causes other cognitive responses: the desire to call the sister and give her a piece of your mind, the belief that your friend would like to be cheered up, the desire to comfort your friend, and so on.

Now consider an analogous *fictional* case: you are watching a film in which your favorite character's older sister forgets to call her on their birthday, even though everyone else remembered to do so. Just like the previous case, you adopt beliefs about both your friend and her sister. This in turn causes you to feel great annoyance and frustration towards the older sister and pity for your favorite character. We can even say, like in the *American Psycho* example from chapter 1, that the filmmakers created a story that is perfectly similar to your real-life best friend so that the content of your emotional state is roughly the same. We can stipulate that the content of the story and our beliefs about it are similar enough that they do not significantly impact our emotions' in inferential role—except, of course, the belief that the fictional characters are *not real*. This belief makes a difference in the emotional response we have towards the characters. It requires that we take the fictional stance towards the object of our emotion. This means that there is a difference in the intentional content between our beliefs about the real-life and fictional cases.

Our emotions about fictional objects have a different intentional content than those about real things. This accounts for differences in an emotion's inferential role. We do not need to posit the existence of imaginary or make-believe emotions to explain the differences between our emotions about fictions and real-life.

2.3. Genuine feelings

Emotions usually involve a conscious feeling. *Prima facie*, we identify our emotions by how they

feel; sorrow, anger, joy, jealousy, pride, etc. all feel a certain way to us. For our purposes, *feelings* are the qualitative bodily responses that are potentially consciously experienced. Chocolate has a particular conscious taste and red has a specific qualitative look; similarly, emotions have conscious qualitative characters. As William James (1890) noted, feelings put the "emotionality" in the emotion, making it salient and important to our lives. Some theories of emotions—call them *feeling theories*—hold that emotional feelings are necessarily conscious; every time we experience an emotion, we feel it (Block, 1995, Goldie 2004, James 1890, LeDoux 1996). According to some views, however, emotional feelings need not be conscious (see, Prinz 2004a, Berridge & Winkielman 2003, Rosenthal 2008). This distinction is not crucial for our current discussion, but it will be relevant in the proceeding sections and chapters when we discuss our actual emotional responses to fictions.

We seem to have genuine emotional feelings towards fictions. The DAV challenge concerns whether these feelings are somehow *unreal*. It is worth noting that even Walton—who, along with Currie, is perhaps the most ardent supporter of the DAV—does not deny that our make-believe emotions towards fictions elicit genuine feelings. He only denies that those feelings are necessary components of the emotion (Walton 1990).

One way to support the claim that our emotional feelings towards fictions are non-genuine would be to show that these feelings are merely illusions. Call this the *illusory feeling thesis*. Consider two other bodily feeling illusions: phantom limb pain (PLP) and the rubber hand illusion (RHI). A phantom limb can be defined as "the continuous awareness of a (or part of a) non-existing or deafferented body part with specific form, weight, or range of motion" (Richardson 2009, 137). There are records of phantoms corresponding to virtually every body part, including arms, legs, teeth, tongue, breasts, bladder, etc. The phantom limb is *embodied* and *felt*. The patient

has the sense that the missing limb remains a part of her body and she may also have sensations in the missing limb. The patient might experience the limb as moving (kinetic illusion) or in a particular orientation (kinesthetic illusion). The patient may experience PLP—pins and needles, itches, dull pains, or acute pains—in the missing limb. While there is no universally agreed upon cognitive or neurological mechanisms associated with phantom limb, potential causes include cortical reorganization or pain memory after the limb is lost (*ibid,* 141).

The RHI is another bodily feeling illusion. In a study by Botvinick and Cohen (1998), a subject sits in front of a table with one hand hidden behind a screen and the other under the table. A rubber hand is placed on the table directly in from of her. If both the real hand under the table and the rubber hand are stroked synchronously by a brush, the subject may feel the touch of the brush on the *rubber* hand, not her real hand under the table. In some cases, subjects might even have the impression that the rubber hand is their own. When this occurs, the subjects will flinch and show a stronger than normal skin conductance response when the rubber hand is hit with a hammer (de Vignemont 2007).

Here we have two examples of bodily feeling illusions. They each result in a feeling in the subject's body part or an inanimate object that does not belong to them—even though the subject *knows* that the limb/hand does not belong to them. Perhaps one could argue that our feelings towards fictions are illusory in a similar fashion. Our emotional feelings towards fictions are not genuine feelings, but illusory, just as the feelings of the phantom limb and the brushstroke on "our" rubber hand are illusory. We feel like we experience some emotional feeling, but we actually do not. It may seem strange that we have illusory emotional feelings—but, then again, PLP and RHI are equally strange to those who have not experienced them firsthand.

Both PLP and RHI engender false beliefs in their subjects. In PLP, the sensations of pain

and movement in the phantom limb cause a false belief that *there is a sensation of pain in my (non-existent, deafferented) limb* when, in fact, the patient lacks is missing the relevant body part. The RHI engenders the false belief that I either a) feel a sensation in the rubber hand, or, b) that there is a sense of embodiment in the rubber hand (the rubber hand is really my own). So both PLP and RHI create false beliefs concerning the origin and locations of sensations in body parts that, in actuality, are not one's own.

Is there a comparable false belief in the case of fictional emotional feelings? I argue that there isn't. When we experience joy, fear, anger, etc. about a fictional object, we "feel" the emotion in the same way we normally would if we felt a genuine emotion about a real object. That is, we experience the emotional feeling as our own, taking place within our actual body, not in a phantom or fake body part. We are not tricked or deluded into having emotional feelings about fictional objects, as in the RHI. There is nothing wrong with our brains, as in PLP. Our emotional feelings are natural reactions to objects that we know do not exist.

2.4. Emotional evaluations

Philosophers, psychologists, and neuroscientists alike all generally accept that emotions involve some kind of judgment. Emotions are evaluative. When we have an emotion, it is because something in our environment—or something that we think, remember, or imagine—bears significance on our lives or the life of someone we care about. This may be a very quick, automatic evaluation, like when we suddenly fear a loud noise behind us or are afraid that we will slip on an unseen staircase. Or the evaluation could be quite complex, like when we experience jealousy towards someone in our workplace. A proponent of the DAV could argue that our feelings towards

fictional objects are not stereotypical states because they are the result of a fictional *evaluation*.

One could argue that we do not have the right kind of evaluative relationship with fictional objects in order for us to make genuine judgments about them. Fictional characters may not be the kind of thing that we can care about, empathize with, feel sympathy for, etc. Again, we *seem* to care about and identify with fictional objects all of the time. We feel very strongly for our favorite television, film, and literary heroes. We want them to succeed and we feel frustrated, sad, or angry when they do not. Still, one could argue that those judgments are not real judgments because the judgment's object does not exist.

This is a big bullet for the DAV theorist to bite. On this view, we cannot genuinely evaluate nonexistent objects. This means that we do not make genuine evaluations of fictional objects, but also of *any other* non-literal or nonexistent object, including hypothetical, imaginary, future, or past objects. This seems intuitively implausible. But since the DAV is ready to deny that we have a genuine *emotions* towards these objects, then they may be willing to deny that these are genuine evaluations as well. We need some independent reason for thinking that the evaluation is genuine.

We need to know whether the object of our emotional evaluation must exist in order for the evaluation to be genuine. It seems like they must, in order for our evaluation to "refer" and not be "empty." My response requires a brief foray into the ontology of fictional objects (see Tullmann & Buckwalter 2014). Specifically, we need to understand the sense of 'exist' found in both the paradox of fiction and similar discussions of emotional responses towards fictions. In other words, *how* do fictional entities exist? This is not a topic that is typically addressed in the paradox literature. In fact, most philosophers take it for granted fictional entities *don't* exist, in a standard sense of the word. But examining this question more closely will help us to see how we can take fictional entities as the objects of genuine emotional evaluations.

Recall my discussion of the ontological implications of the fictional stance. There are at least three ways in which a fictional object exists.[15] First, an existent object might be a concrete particular. This is the sense of exist we use when we say that a comet, person, rock, or painting exists. They are all a part of our actual, corporeal reality. This is clearly not the sense in which fictional objects exist (if they do at all), but it *is* assumed by the paradox. Consider the standard responses to the paradox's second proposition: we do not run screaming from the theater when we see a zombie on a movie screen, so we do not genuinely believe that the zombie exists. Zombies are not actual, physical creatures in our world.

Second, a fictional object *could possibly* exist in another world. In other words, fictional entities might be possible objects in our (actual) world, but actual objects in another possible world. David Lewis (1973), Alvin Plantinga (1974), and Saul Kripke (1980) all held versions of this view, although Kripke has rejected this view in his more recent work (Kripke 2013). There are many problems with the possibilia thesis, as we saw in chapters 1 and 3. Descriptions of fictional entities are incomplete in detail, whereas possible objects are taken to be complete. For example, it is indeterminate whether or not Conan Doyle's Sherlock Holmes had a mole on his back. If Holmes were a possible object, then he either would or he would not (Thomasson 1999). We are also unable to distinguish between possible objects (Quine 1953). Finally, fictional objects are created and causally dependent upon authors and audiences. Even if Sherlock Holmes was a denizen of another possible world, it would not be the same Holmes that Conan Doyle wrote about in his stories (Sainsbury 2010).

[15] I have not discussed Meinongianism here (Meinong 1960, Priest 2005) or the Sartrean sense of imaginary entities (Sartre 1991). See Thomasson 1999 for an overview. One can also deny that fictional entities exist in *any* sense. See my discussion of pretense theories in chapter 2, as well as Quine 1956, Russell 1905, and Sainsbury 2010.

So we should be skeptical of the view that fictional entities exist in this sense of the word. But there is another, more plausible sense in which fictional entities may very well exist. The last sense of 'exist' is the inverse of the first two: a fictional object is neither a concrete particular nor *possibly* a concrete particular. Instead, fictional objects might be abstracta of some kind, as I suggested in chapter 3. They are created and dependent on people and conventions. Fictional entities would go out of existence if everyone forgot about them and there was no record of them. We even, in a way, interact with fictional entities. Discussing a fictional object can be used to justify one's actions and form non-assertoric propositional attitudes (see also Carroll 1990). We can also form thoughts about and respond unconsciously and non-cognitively to fictional characters.

With this last sense of 'exist' in mind, we can turn back to the question of how we can form genuine evaluations of fictional entities. Importantly, this sense of 'exist' should appeal to any theory of the nature of emotions. First, consider feeling theories. On this view, emotions are non-cognitive in the sense that they are not comprised of thoughts or beliefs. Yet emotions are also *not brute.* They are evaluative—they involve judgments of objects in one's internal (mental) or external (physical) environment. Because the feeling theorist claims that emotions are basic in this way, the object of the emotion need not be a concrete particular. I will examine these 'affective appraisals' in more detail in the following sections; for now, it is enough to note that we will respond emotionally to any object—real or unreal—that we perceive to bear on our well-being or that of someone we care about. If this is the case, then we may have genuine emotional evaluations of entities that only exist in the last, minimal sense.

We can give the same sort of explanation for cognitive theories of emotions. Like feeling theories, cognitine theories of emotion hold that our emotions involve an evaluation of an object

in our internal or external environment that bears on our well-being. Walton argues that this emotion is not genuine because we are not motivated to act in the right way. We have already seen, however, how we can eliminate this kind of functional argument about emotions. It is perfectly possible to be motivated to act based on our emotional responses towards fictions.

Consider an example:

> The unlucky-in-love Becky [thinks about] the *character* Mr. Darcy, just as he is described in Jane Austen's novel. It is this Mr. Darcy that serves as the object of her emotions (longing, wistfulness—but also, this time, a strange jealousy for Elizabeth Bennett and regret that such a man doesn't exist). These emotions may even cause Becky to engage in certain peculiar actions: she places her copy of *Pride and Prejudice* on her nightstand in an oddly affectionate manner and makes caustic remarks about Elizabeth Bennett's contrariness to her friends (Tullmann & Buckwalter 2014, 792).

Becky emotionally responds to a fictional entity, forming evaluations of the characters Mr. Darcy and Elizabeth Bennett; for example, that she and Darcy would make a good couple and that Elizabeth is a threat and is clearly up to no good. As strange as this example may seem, it shows how we can be motivated to act based on our judgments of fictional entities—not just general actions caused by our beliefs about fictional characters, or to people that we think are like those characters, but actually based on our beliefs concerning fictional entities themselves. Mr. Darcy is neither a concrete or possible object. If he exists at all, it is in the sense of being an abstract object. If we can have genuine judgments of Mr. Darcy in this sense, then there does not seem to be a strong argument against the idea that we can form genuine emotional evaluations of any fictional entity.

I have considered, and rejected, three arguments against the SAV of emotions, and will return to a fourth (about the semantics and normativity of emotions) in §4. My arguments should apply to any theory of the nature of emotions. Emotions could be beliefs, judgments, feelings, or

some other kind of appraisal. Regardless of what constitutes an emotion, one would have to show that it is that feature which makes the emotion non-genuine in order to show that the emotion *itself* is not genuine. The best way to do this is to argue that the emotion does not lead to the right kind of behavioral output. And this argument fails. Importantly for our purposes, my conclusion makes it easy for us to sidestep the traditional talk of the paradox of fiction altogether.

3. An appraisal theory of emotions

The purpose of the previous section was to dissolve the paradox of fiction and, in so doing, motivate the SAV of emotions. I did not rely on a particular theory of emotions to do this. However, there remain several questions concerning our emotional responses towards fictions that *do* require some theoretical commitment: What sorts of mental states are involved in our emotional responses to fictions? Are we always aware of our emotions toward fictions? Why do we consider fictional entities to be worthy objects of our emotions? For that matter, how do we emotionally respond to fictions in the first place? Of course, similar questions arise when dealing with real-life emotions. Indeed, these are important questions faced by any general theory of emotions, not only theories of fiction.

So far, I have treated emotions as a kind of evaluation without specifying the nature of that evaluation. In this section I will try to get clearer on what I mean by this. I argue that emotions are constituted by multi-level appraisals. Emotions are not brute reactions to stimuli. While I hold that emotions are comprised of non-cognitive content, they can be understood as cognitive in some

other sense: they are intentional (about something), they are evaluative, and they carry information.

While I think that our emotional responses are constituted by multi-level appraisals, I want to stress that a proponent of the SAV may adopt virtually any theory of emotions. I favor an appraisal theory because it can explain the interesting and difficult features of our emotional experiences, both about fictional and actual things. It is not my goal here to fully motivate my multi-level appraisal theory, or to fend off every potential argument against it. But, as we will see, I do think that my theory may be the best for explaining a moral psychology of fictions.

3.1. Fictional objects as emotional objects

Emotions play a strong motivational role in our lives. They help guide our actions, shape our relationships, and inform our decisions. Appraisal theories attempt to capture the motivational aspects of emotions. We can think of emotional appraisals in terms of biological fitness. An organism constantly evaluates its environment for features that might bear on its survival and flourishing. Emotions clue us into these features, both consciously in terms of emotional feelings, and unconsciously in terms of behavioral priming and automatic physiological responses.

I will discuss three important features of appraisal theories before turning to my own account of emotions: the content of our emotional appraisals, the causal and constitutive components of our emotions, and whether an emotion is best understood as a process or discrete mental state. Finally, I will address two general objections to multi-level appraisal theories that can be found in the emotion literature.

First, appraisal theories of emotions need some story about how a subject evaluates an object. For my purposes, this story must also apply to fictional objects. It is easy to see how a real

life object, like a snarling dog or a person with a weapon, is a worthy emotional object. It is less clear how and why a fictional object would be, since the fictional object does not bear on our well-being in the same way as the real life object does.

A traditional way of understanding the intentionality of emotions—what an emotion is *about*—relies on the cognitive stance that emotions are comprised of propositional attitudes. According to a belief-based theory, our emotion is constituted by the belief that *X*. Other cognitive theories could replace the belief state with a thought or judgment that *X*. This could explain how fictions could be the objects of emotions; we make a judgment that a fictional villain is evil or terrifying, believe that the villain is trying to harm the protagonist, or merely have the thought that the villain is up to something no good. In turn, these cognitive states may lead to feelings of fear and apprehension.

Unfortunately, this approach is not available to any theory that denies that the initial appraisal of the stimuli is non-cognitive and non-propositional (as mine will). So before we can understand how fictions can be the objects of our emotions, we must understand how non-cognitive theories of emotion explain intentionality. Two proposals include Richard Lazarus's "core relational themes," (Lazarus 1991, Smith and Lazarus 2000), a concept also adopted by Jesse Prinz (2004a & b) and Peter Goldie's "emotion-proper properties" (2004). Goldie describes these as the property that things must possess in order to cause a certain bodily state associated with a particular emotion (Goldie 2004). Core relational themes work similarly. According to Lazarus, we respond emotionally to certain stimuli because those stimuli represent certain themes. For example, we feel sad when we encounter a situation in which we experience a *loss* of some kind. We fear a snake because it represents the core relational theme of causing *harm* to us or one we care about. The same applies to other emotions, including complex emotions like jealousy: romantic jealousy

arises when we perceive or imagine that that one's significant other has been unfaithful (Prinz 2004a).

Neither the emotion-proper property nor the core relational theme requires that we have an assertoric belief about the object, situation, or state of affairs the emotion is about. They do require us to be able to perceive emotion properties, or at least be able to make low level judgments about our perception of emotional expressions. We can also be seriously wrong about what causes an emotion. We might, alone in our dark bedroom, think that a quick-moving shadow is a burglar attempting to break into our room, when in fact it is was caused by a car passing outside our window. Just the perception of this shadow may be enough to set off an emotional state. But our fear does not require any thoughts or beliefs about what this shadow actually is or the potential harm it can cause. This is the key to understanding how we can have genuine emotions about fictions. We associate objects on a screen, or in a play, or even imagistically (as when reading a novel) with an emotionally relevant property.

There is one last point we need to emphasize here. An emotional appraisal is generally about the subject's own well-being. But there are obvious counterexamples to this. Our emotional responses often concern other people: we may feel pride on behalf of a family member's achievement, anger at an injustice committed against a friend, or joy at a partner's success. None of these emotions involve appraisals that directly involve applying a core relational theme to an object in terms of one's own well-being.

Most theories of emotion extend the notion of 'well-being' to include those people and objects that we care for and about. I doubt that we need to think of others as an extension of ourselves in order to emotionally respond on their behalf. Nor must we empathize or identify with them. I discussed this in the previous chapter; empathy is not necessarily the basis of our emotional

responses about fictional objects. Instead, I think that concern for another might be enough for us to emotionally respond on their behalf. We must make some kind of positive evaluation of the person or object. This is also why we feel anger, sorrow, joy, hope, pride, embarrassment, shame, etc. on behalf of fictional characters: we care about them.

As we have seen, many theories argue that emotions either involve cognitive, propositional attitudes such as beliefs, thoughts, judgments, or non-cognitive states like feelings or perceptions of bodily changes. In general, appraisal theorists occupy a middle ground between the cognitive/non-cognitive extremes. The challenge is to determine *which* psychological processes are involved in our emotional appraisals.

3.2. Appraisal theories

Different appraisal theorists have different ideas about what constitutes an emotional appraisal. However, there are some features that all appraisal theories share, besides the fact that they all emphasize a subject-environment relationship (see Moors et al 2013). First, appraisal theorists generally deny that propositional attitudes are sufficient for emotions, even if they are necessary for them (Arnold 1960, Ellsworth 1994, Roseman 1996, Smith & Kirby 2001, Scherer 1984, Smith & Lazarus 1990, Stein et al 2000). So appraisal theories may escape the brunt of the evidence from neuroscience and psychology that has recently come out against belief-based theories of emotion (Prinz 2004a). However, the appraisal theorist may still run into trouble in this regard if they hold that a cognitive appraisal precedes and causes sub-personal bodily reactions to emotional stimuli (Scherer 1984). This does not mean that the appraisal must be deliberate. We do not have to consciously evaluate an object. Instead, the appraisal may be fast, automatic, and "virtually

instantaneous" (Smith & Lazarus 1990). This leads to my next point. Appraisal theorists also deny that emotions are necessarily conscious. This makes sense; we do not always think of ourselves as consciously evaluating objects in our environment.

Third, each appraisal theorist posits some kind of low level monitoring system that tracks emotionally relevant properties and objects in one's environment. The monitoring system may track objects in one's environment (or objects of a thought one considers) that are relevant to one's goals, motives, or well-being (Scherer 1984, Roseman & Smith 2001, Smith & Lazarus 1990, Stein et al 2000). Positive emotions like joy and pride arise when the object is appraised as goal-congruent, promoting, or compatible. Negative emotions like anger or shame interpret their object as goal-incongruent.

Finally, appraisal theorists agree that we can individuate emotions in terms of appraisals. This is a problem that has plagued noncognitive theories of emotions: it does not seem like each emotion has a unique qualitative feel or set of bodily reactions. Appraisal theories accept this. What makes an emotion a token of anger, indignation, rage, or annoyance is the particular appraisal that one makes with respect to an emotionally relevant property or theme. This leaves room for emotional valence—the positive or negative qualitative feel of emotions—to also play a role in emotion individuation. But valence alone is not fine-grained enough to determine the difference whether a particular emotion is anger or indignation, sorrow or grief, joy or elation.

It is sometimes unclear whether an appraisal *constitutes* an emotion or merely *causes* it. Scherer (1984) seems to argue that appraisals *elicits* the emotion. The nature of the emotion itself remains a mystery. Paul Griffiths (2002, 2004) makes this point in his extensive work on the different types of emotions. To be fair, Scherer does point out that the causal/constituent question depends on what one means by an 'appraisal.' Part of a solution here will depend on whether one

thinks that emotions are an extended process or a discrete mental state, which I will address momentarily.

Appraisal theorists agree that emotions involve the ability to perceive, mentally represent, and process sensory information and other occurrent mental states. My multi-level appraisal theory will draw specific lines concerning each of these aspects of the emotional appraisal. I think that appraisals *constitute* an emotion, rather than merely *cause* them. I agree that emotions are not necessarily conscious, are generally non-propositional, and require some kind of monitoring device.

Appraisal models also differ with respect to whether emotions are best understood as continuous processes or discrete, categorical states. Again, most appraisal theorists think that we continuously subconsciously monitor our environment for significant stimuli. The main controversy within appraisal theory concerns whether there is a set structural procedure that each emotion must necessarily follow in order to be classified as a particular emotion or whether the process is more flexible and open ended. Here I will examine one prominent process-oriented model from Richard Lazarus and C.A. Smith's (Smith & Lazarus 1990).

Lazarus and Smith's argue that the appraisal of a core relational theme may be automatic and unconscious (*ibid,* 629). There are two basic parts of the emotional appraisal. The first is schematic: the subject appraises a stimulus in terms of how it bears on its well-being, on the basis of past experiences with similar encounters. This requires that associative memory networks are activated, drawing upon relevant information about the object and past responses to it. The same stimulus information is sent for cognitive conceptual processing. This involves more abstract, deliberate, and conscious cognitive processes through which the subject is able to evaluate the adaptational significance of the stimulus more actively and accurately (*ibid,* 630).

Importantly, Lazarus and Smith do not think that there is a set sequence or structure specifying how the stimulus information is processed. Conceptual processing will generally follow the initial schematic processing due to the nature of the neurological components involved. But the conceptual processing will also be available for *further* schematic processing. This may result in multiple feedback loops in which subconscious schematic processing impacts and modifies conceptual processing and vice versa.

I am highly sympathetic to Lazarus and Smith's process-oriented emotional model, although their theory has come under scrutiny for being empirically implausible. Like them, I will argue that emotions are temporally extended, involving multiple different appraisals almost at once, each impacting the other in a feedback loops (see Damasio 1994). However, we can still isolate emotions in terms of discrete states. I want to suggest that the major emotion categories (joy, sorrow, anger, pride, curiosity, etc.) are "somewhat crude attempts" to describe our emotional experiences (Scherer 1984, quoted in Roseman & Smith 2001, 14). In fact, I think that our emotions are, to a certain extent, *epistemically indeterminate*. This means that we do not always have access to what causes a particular appraisal or how stimulus information is processed. It may be a matter of interpretation of one's environment and internal milieu that determines which emotion one currently experiences, or experienced in the past.

I will not spend a great deal of time defending my appraisal theory against its foes.[16] However, there are two general objections to traditional appraisal theories that I wish to eliminate right away. There has been a general trend in recent years towards theories of emotions that emphasize bodily (somatic, phenomenal) features of emotions. One objection against appraisal

[16] See the Appendix at the end of this chapter for a discussion of four more potential objections to my multi-level appraisal theory.

theories is that they are *too* cognitive. Traditional appraisal theories hold that an emotion must be caused by or constituted by a cognitive judgment of something in our environment, where 'cognitive' here means more than simply possessing an intentional object (see Prinz 2004a). In light of my discussion here, however, we can see that this objection does not necessarily hold. Most appraisal theorists argue that the appraisal that constitutes or causes the emotion can be fast, automatic, unconscious, and non-propositional. They also generally leave room for later cognitive processing. Most appraisal theories discuss the role of cognitive processes in our emotions, but they do not posit that cognitive or propositional evaluations *are* the emotion.

Griffiths (1997, 2004) raises the objection that appraisal theories are "ecological," primarily concerned with explaining the significance of the environment for an organism. In doing so, such theories do not adequately concern emotions as mental states and "[abstract] away from the details of any particular psychological process" (Griffiths 2002). By emphasizing the way in which emotions are brought about via core relational themes or emotion-proper properties, appraisal theories ignore the cognitive and subpersonal mechanisms that are involved in actually bringing about a discrete emotion. This objection also strikes me as missing the point of appraisal theories. Griffiths is correct in thinking that appraisal theories emphasize the conditions that bring about the emotion, but, each appraisal theorist also generally has a story about the psychological processes and neural mechanisms involved in discrete emotional states (see, in particular, Laxarus 1991 and Scherer 1984).

I will now turn to a discussion of my own multi-level appraisal theory which, I argue, adequately surmounts these two objections.

3.2. A multi-level appraisal theory

I want to propose a multi-level appraisal theory that is similar to Lazarus and Smith's process model, but differs with respect to the nature of those appraisals and how we individuate specific emotions. I will also draw on Scherer's point that emotion individuation is largely a matter of interpretation.

I suggested that we can understand appraisal theories along three different dimensions: the nature of the appraisal itself, the components of the emotion, and whether the emotion is structural or process-oriented. I will present my own view along the same guidelines. I agree with the other appraisal theorists that we should understand emotions as responses to objects possessing emotionally relevant properties. However, in light of Griffith's ecological objection, I want to be more specific about the psychological and neurological mechanisms that may be involved in a subject's recognition of these properties.

Following Lazarus, I argue that emotions generally involve two different appraisals. The first arises when a stimulus in one's external (physical) or internal (mental) environment causes a mental representation of something that bears on one's well-being. The mental representation comprises the emotive value of the stimulus, which may be learned or innate. For example, humans may have an innate predisposition to fear heights or looming objects and a learned fear of monsters under the bed, certain people or social interactions (Damasio 1994). When we perceive, imagine, or think about the looming object or monster (or something like it), we form a mental representation that triggers an emotional response.

I understand this initial representation as a kind of appraisal that is subconsciously and subpersonally implemented. We respond to perceived objects (in our environment, or intentional

objects of mental attitudes) as pleasant or unpleasant, or as things to avoid or approach. Information about the stimulus is transferred to areas of the brain that are involved in encoding emotional reactions, such as the sensory thalamus, the affective division of the striatum, the orbitofrontal cortex and, finally, the amygdala (LeDoux 1996, Mello & Villares 1997, Rolls 2000; see Schroeder & Matheson 2006 for an overview), each of which are involved in initiating bodily reactions and behaviors, such as an increased heart rate or freezing response.

At the same time, the sensory stimulus information takes a slower track to cortical regions of the brain for further processing and input from cognitive faculties. LeDoux (1996 & 2012, also LeDoux & Phelps 2000) call the initial pathway the "low-road" of emotions, and the slower pathway the "high-road." The sensory information involved in the initial appraisal reaches the amygdala or hippocampus after first being processed by the sensory thalamus. The same information is processed by the sensory thalamus, and also sent to the sensory cortex for cognitive processing and availability for consciousness. Here, the subject's knowledge about the stimulus, her beliefs about her current environment, and desires concerning her future goals may all influence how the stimulus information is processed. She may also deliberate, draw inferences, and make decisions concerning the potential value of the object.

Let's say that a subject perceives a potential threat in her environment: something she finds fearful. Any number of objects could possess a "fearful" property, from a snake that crosses our path to an important test that looms in the near future. Our subject perceives the snakes, drawing on information from long-term memory that allows the subject to appraise this particular specimen. The value of this appraisal may vary from person to person: our subject may be quite fond of snakes and consider them fascinating, beautiful creatures. Her stored information of snakes may include these features, whereas another person might associate snakes with harm or icky-ness. The

stimulus information is also sent via the cortical pathway for cognitive processing. Our subject's beliefs about her surroundings may correspond or disagree with her initial appraisal of the snake. She might be particularly fond of this breed of snake, or, alternatively recognize that this is not a pet snake and so should be avoided.

The multi-level appraisal theory posits that there are generally two different appraisals involved in our emotions, one precognitive and one cognitive (call them *affective* and *cognitive* appraisals, respectively). But emotional processing may not stop there. Once the stimulus information is processed by different cortical regions, it may be sent *back* to the low level precortical regions, such as the amygdala and hippocampus. Several different things can happen there. The initial reactions resulting from the initial affective appraisal may be extended, intensified, modified, regulated, or even eliminated in light of information from cognitive processing.

This forms a feedback loop in which subcortical information processing influences cognitive processes, and vice versa (see Damasio 1994 and Smith & Lazarus 1990). This may occur over the course of a second or two, or perhaps even longer as more information about one's environment is evaluated and weighed with respect to our affective and cognitive appraisals. All the while, one's perceptual systems monitor the environment for relevant objects or information which may further impact one's affective processing. I do not argue that there is a set, structured method in which different appraisals follow one after another, resulting in a specific emotional state. My view is similar to Smith and Lazarus's process-oriented account, in which schematic and conceptual processing occurs continuously.

In some cases, the cause and object of our emotion may not be known to us or we may misattribute it (as in the famous Schacter and Singer studies, 1962). I would argue that these

emotions are *epistemically indeterminate*; we are unaware of or lack access to the causes and/or object of our affective responses such that we cannot accurately identify what emotional or affective state we are (or were) in. Determining the identity our emotional state may be post hoc, like when we say that a film is "happy" or "sad" even though it, in fact, involved a wide variety of emotional appraisals and responses.

Scherer (1984) argues that identifying particular emotions is largely a matter of interpretation. I agree. Like other appraisal theorists, I argue that we respond to stimuli that correspond to an emotionally relevant property. But which objects correspond to which property will vary from person to person; each subject will interpret different things as fearful, anger-inducing, or joyous. Moreover, whether we identify an emotion as anger or rage, sorrow or grief, shame or guilt largely depends on contextual information and our relationship to the emotional object. For example, I might describe my downcast feelings, thoughts, and bodily responses as embarrassment or shame depending on what sort of norm I have violated. So we not only identify emotions in general in terms of appraisals, we also identify our own emotions in a similar way.

Note that I have not argued for particular necessary conditions that are jointly sufficient for an emotion. My position leaves open the possibility that an emotion involves only one appraisal, affective or cognitive. This implies that there might be completely "cool" emotions that do not result in bodily reactions of any sort. This may be theoretically possible. If stimulus information proceeds along two pathways, it is possible that the lower pathway becomes blocked while the cortical pathway remains intact (but disconnected from the amygdala and other brain areas that lead to physiological responses; see Figure 5.1). Some variation on this idea would help to explain cases of flattened affect in which we see cognitive processing with lessened non-cognitive processing, as well as phobias—non-cognitive processing without cognitive processing (see

Roseman and Smith 2000). However, I do not think that many such cases will occur in everyday emotional responses. Even phobic or flattened affect cases will involve some kind of affective and cognitive processing, even if they are irrational or inconsistent thoughts or beliefs. And the mere thought of an emotion-laden stimulus may result in some bodily reactions—such as a galvanic skin response, increased pulse, or dilated pupils—even if these reactions are not consciously *felt*.

This means that the traditional bifurcation of cognitive versus noncognitive emotion theories may be largely defunct. It should be clear that, on my view, both noncognitive and cognitive processes play valuable roles in bringing about one's overall emotional state. This theory is non-cognitive in the sense that emotions are not caused by an initial belief, judgment, or thought, but rather by perceptual processes. Both the initial and later appraisal may occur quite rapidly and without conscious awareness or feeling. However, emotions generally involve cognitive information processing of environmental or internal stimuli. This means that emotions are intentional; they are about something.

Finally, I argue that a multi-level appraisal theory of emotions is well-suited for bolstering the SAV. My comments against the DAV functional argument strongly rely on the notion of emotional regulation and modification. It is possible that we are initially motivated to act or respond emotionally to fictional stimuli when we are engaged with a fiction, but due to other cognitive judgments, beliefs, etc., we do not actually follow through with an action. This process can now be characterized in terms of the multi-level appraisal view. The regulation and modification of the emotion occurs with the later, cognitive appraisal which, in a sense, blocks any action that we might be primed to take. Of course, neither appraisal need be conscious, so this is consistent with our actual experiences of engaging with fictions.

4. Rational fictional emotions

There is one more hurdle to scale before we have clearly established the SAV for emotions: explaining emotional rationality. This follows from the idea that emotions are evaluative; we tend to think that our emotions can get things right about the world. Our anger, fear, sorrow, etc. pick out some feature in our environment, and they do so correctly. Explaining emotional rationality is a troublesome business in its own right. The challenge is propounded with *fictional* emotions. But how could it be that our emotions concerning fictions are rational in the way that our real-life emotions often are? If we cannot develop some account of rational fictional emotions, then we can perhaps say that these emotional responses are distinct in this way.

I will argue for an account of rational fictional emotions in this section. There are several different ways in which we can characterize this. Here I borrow concepts and terms from two sources: Ronald de Sousa's work on the semantics of emotions (their truth-conditions) (2002, 2004) and Justin D'Arms and Daniel Jacobson's extensive work on emotional normativity, norms concerning when and how we should experience emotions (2000a, 2000b, 2008).

There are several ways in which emotions can be understood as being rational. First, an emotional response may *fit* its object. An emotion fits its object if we have some reason to feel it. We can compare the fit between an emotion and its object to that of a true belief and a state of the world. Both spiders and battlefields may be fitting objects of fear; this evaluation is apt in some way, as being proper formal objects. Our colleague's promotion may be a fitting object of jealousy. An off-color joke may be a fitting object of amusement. On my view, emotions fit their object in case we have some reason to have them for that particular object (compare this to D'Arms and

Jacobson's slight different account of fit, which they characterize in terms of a response-dependent feature of the object that does not require reasons or norms).

Second, emotions may have truth values. De Sousa explains emotional truth in terms of the *success* and *satisfaction* conditions. An emotion's success corresponds how they describe their object; there is some*thing* that bears the relevant emotional property. In order for an emotion to be satisfied, however, implies that the emotional property actually obtains in some object.

Finally, we can also speak of an emotion's *propriety*. I take propriety to carry morally normative implications; it suggests that there are appropriate contexts in which we can or should have certain emotions. Importantly, fit, truth and propriety may not always match in any particular object. An emotion may fit its object, or even be successful in de Sousa's terms, but not necessarily be the proper response to take. For example, even if a battlefield is a fitting object of fear, it may not be *proper* for a soldier to feel if he or she has an important task to fulfill. If our colleague is also our friend it may be improper for us to be jealous of her promotion—we *should* be happy for her—even if it is fitting for us to be, since, perhaps, we were also due for a promotion and did not get one. When we conflate the fit and propriety of emotions, we commit *the moralistic fallacy*: taking the morally normative implications to be built into our emotional responses towards things in our environment (D'Arms & Jacobson 2001a).

I will discuss each type of emotional rationality in the course of the following sections, focusing particularly on the special issues that fictions present. Ultimately, I will argue that we should characterize the rationality of our emotional responses in terms of fit (apt reasons we have for justifying a particular emotion) rather than truth (success conditions).

4.1. Is it epistemically rational to have emotions about fictional objects?

Since Plato and the Stoics, many philosophers have held that some emotions are epistemically harmful and thwart our ability to reason clearly. However, recent literature in cognitive science has begun to debunk the traditional bifurcation between rationality and emotions, showing that emotions are often necessary (or at least useful) for planning, making important decisions, and making moral judgments (Ben-Ze'ev 2001, Damasio 1994, Gordon 1987, Nichols 2004, Solomon 1993, etc.). It is debatable that these benefits extend to our emotions about fictions. For instance, Currie (1995) points out that the concept of epistemic emotional rationality is puzzling because fictions generally present a reader with a great deal of false information. This could undermine her ability to function in the real world if she takes it literally.

We have already encountered Radford's argument that fictions commit us to inconsistent beliefs. Most theorists now reject the notion that emotions necessarily involve beliefs, especially the belief that the emotional object exists. According to the multi-level appraisal theory that I have presented, we can experience emotional responses towards fictional objects even if we know that these objects do not exist in our world. We can eliminate Radford's worry that fictional emotions involve us in a rational inconsistency by denying that emotions require a belief in the existence of the object of the emotion. However, one might still worry that our emotional involvement with fictions somehow hinders our everyday epistemic capacities, since, as Currie points out, fictions do not present us with factual information about the world. If our goal is to gain new information and align our thoughts with the truth, then why are we wasting our time with fictions? Interestingly, though, we also tend to think that we can learn something valuable from fictions. Fictions would be epistemically useful in this sense.

Of course, we are not always concerned with gaining factual information from ficitons. We

read novels and watch movies for entertainment, to relax, and to escape our everyday lives. We may not expect to acquire factual truths from fictions. Interestingly, many philosophers argue that fictions have something to teach us even in these capacities. Instead of factual truths about the world, we learn about emotional, moral, and psychological tendencies and trends, while also honing our own moral and emotional capacities.

Our emotional responses to the characters in a fictional often clue us in to significant features of a work. Jenefer Robinson (2005) argues that we can may gain a *sentimental* education by engaging with fictions. When reading a novel or watching a film, we witness the exploits of a character, access his thoughts, and analyze his behavior. Robinson claims that fictions can teach us important and interesting folk psychological facts if we are aware of and reflect on our emotional reactions to a character. We may be surprised by the emotions we experience in the course of a story and how our opinion of a character shifts as the plot progresses. We can learn about our own emotional preconceptions upon later reflection of our emotional experience. Robinson also suggests that our actual experience of reading a novel helps our sentimental education. We learn to focus our attention on various details about a person or character, shift our points of view between different characters, and reflection on our responses.

I agree that the emotional content of fictions can be highly useful for informal emotional training. I would also like to add the idea that emotions provide meaning and context to the fictional scenarios. Robert Gordon (1987) argues that we are often unable to correctly interpret a scene if we are unable to understand the emotional expressions and displays of the people in it. I argued in the previous chapter that we understand another's behaviors, decisions, and mental states partly by understanding their emotional reactions. Gordon's point is that some fictions would make less sense to us without our *own* emotional responses toward them.

Consider the Heider and Simmel movie in which several triangles move about a screen and the participant in the study is invited to narrate what takes place. Participants generally describe the scene as if the objects moving about had mental states, referring to the triangles as "he" or "she" and interpreting their behaviors in terms of the triangles' desires and emotions. This exam is utilized to test the participant's social cognitive capacities, their ability to recognize and interpret the mental states of others (Baron-Cohen 1990, Frith & Happé 1994). It may be that the participant recognizes emotional behavior in the objects and, in turn, experiences emotions themselves. The participant may fear for a particular triangle, feel anger toward another, and be happy and relieved when the big triangle is thwarted in the end of the clip. These emotions are the basis of our feeling with fictional characters, as well as people in our real lives. Emotional responses illustrate who and what we find important and imbue a situation with significance.

A study by Heberlein and Adolphs (2004) further suggests that we project meaning onto the world via our emotions. In their study, they examined the ability of patient SM, who has complete, bilateral damage to the amygdala (which is generally associated with processing emotional information) due to Urbach-Wiethe disease. SM has normal visual perception, attention, language ability, and IQ, but seems to be impaired in making some emotional and social judgments about human faces. In the study, SM was compared with nine other healthy controls (matched for age, gender, and education) as well as subjects with damage to the OFC (orbito-frontal cortex, also associated with emotional and social processing) in her responses to the Heider and Simmel film. This test was specifically designed to examine the role of the amygdala in influencing social meaning, ruling out other factors such as intelligence or damage to other parts of the brain that are associated with emotions. Heberlein and Adolphs found that, in comparison to both control groups, SM described the movie in less emotional and socially significant terms. SM discussed the

movement of the shapes in the film instead of using social and emotional descriptions; for instance, saying that "the big triangle moves inside the square" instead of "the bigger triangle was in control…he wanted to destroy things."

In sum, our emotional engagements with fictions may have practical and educational value even if fictions are not truth-seeking. Our emotional responses towards fictions are epistemically rational in the sense that they help us to develop cognitive skills that are useful in real-life cognitive processes.

4.2. True emotions

One way of understanding the rationality of emotions is in terms of their potential truth-aptness. Consider how the multi-level appraisal theory understands emotional appraisals: we perceive some object in our environment as bearing on our well-being (or that of someone we care about) because it embodies an emotionally relevant property. The fact that emotions have truth conditions would suggest that the emotional property must actually correspond to the object it describes. De Sousa (2002, 2004) has developed an account of emotional truth that may shed light on emotional truth. According to his schema,

> E(*p*) is *satisfied* iff *p* is true;
>
> E(*p*) is *successful* iff *p* actually fits E's formal object,

in which *E* is the emotion and *p* is the corresponding state of affairs (de Sousa 2004, 170).

Suppose that the formal object of fear is "the dangerous." It is appropriate to feel fear towards objects that we evaluate as dangerous to us or someone we care about. An emotion is

satisfied if it *refers* to an object correctly; it is successful if it *evaluates* its object correctly. Consider the fear of falling off the side of a cliff when you are on a hike. This fear is *satisfied* if you recognize that there is some object (the cliff) that is the proper object for fear. Perhaps we are quite close to the cliff so that it is a real possibility that we could take a tumble and doing so would be harmful. Our fear is only *successful* if that evaluation is correct. If it is likely that we could trip and fall off the side of the cliff, then our fear is successful. But we can imagine a situation in which it is very unlikely that we would fall because there are guard rails along the cliff edge or our path never comes within fifteen feet of the edge. In that case, our fear would be satisfied, but unsuccessful.

Could we have a successful, but unsatisfied fear? Since satisfaction is a reference relation, de Sousa argues that it requires that there is some actual object that is the target of one's emotions. It is unclear what sort of ontological status the object must have (concrete, possible, or abstract). Could the fear of a monster be satisfied? De Sousa doesn't specifically address this issue; however, he does refer to the fear of a monster as *unsatisfied*, meaning that the fearful object doesn't exist. Consider my discussion of 'existence' in §2.3. I argued that a fictional entity may exist as an abstract entity. If that is the case, then it would seem that a fictional object *could* be satisfied because it does actually exist and is represented in a certain way (as having emotional properties). It just isn't a concrete particular. Furthermore, fear of a monster may still be successful even if we do not recognize that it exists. This is so if we recognize that the fictional monster is represented as a dangerous creature.

Even if the emotion is unsatisfied, it could still be true on this schema. De Sousa argues that emotional truth only requires *success.* When successful, we have made an appropriate evaluation of the formal object of the emotion even if it does not map on to an actual, concrete

object. This is perfectly compatible with my multi-level appraisal theory. I have argued that the initial affective appraisal of a fictional object, like a monster, may occur in spite of one's knowledge that the monster is not real. This is because the recognition of the potentially dangerous stimuli automatically results in an affective appraisal as if the monster *were* real. It isn't until a later, cognitive appraisal that we evaluate the object of our fear in terms of its existence (see Harris 2000). We may initially experience all the tell-tale signs of fear as though the monster were real: sweaty palms, covering our eyes, racing heart, etc. Slower cognitive feedback influences, controls, or regulates the initial appraisal and its behavioral responses. Our initial affective appraisal of the object of the emotions is similar to de Sousa's success condition: it evaluates a stimulus as dangerous even if it is not real. The second, cognitive appraisal works like the satisfaction condition, evaluating whether or not there actually is an object to be feared, and regulating the emotion accordingly.

4.3. Truth vs. fit

Emotional truth is based on whether the emotion appropriately describes its object; whether our emotion evaluates its object *successfully*. In order to see how an emotion can be true, we must also explain how it can be *false*. One might think that emotions are completely subjective and vary from person to person. Can we have a relative notion of emotional truth?

The challenge for this view is to explain how the emotional success is not merely, or completely, subjective. Emotional truth cannot be like (gustatory) taste, in which there is no way to determine whether a taste is incorrect, because taste *cannot be* incorrect. Taste is subjective; you either like ice cream or you do not. We cannot claim that my dislike of ice cream is false.

De Sousa saves his conception of emotional truth from complete subjectivism by arguing that an emotion's success does not merely depend on an individual's idiosyncratic evaluation of an object. Three factors are involved in evaluating an emotion's success: an individual's personality and background values, biology, and society/culture. According to de Sousa, only the first of these lends itself to a subjectivist reading. First, how one responds to a perceived act of injustice depends on one's personal context, history, beliefs, and values. Second, biological factors, such as innate or basic core relational themes or emotion proper properties, will likely play a role in the correctness of an emotion. For example, it may be that we are innately predisposed to respond with anger towards certain acts that we witness (e.g. violence). However, biology alone cannot tell us how to successfully emotionally respond to certain objects. Social and cultural norms will set a standard of truth as well. Every culture may associate anger with perceived harm and sorrow with loss, but different cultures with count different instances and circumstances as examples of harm and loss. The responses of our collective social group will also factor into a standard of emotional truth.

De Sousa leaves his notion of truth purposefully open ended, with no set rules for determining truth in any particular situation. He does not see this as a problem for his account. However, even with the various personal, biological, social and cultural norms in play, it is likely that the truth of an emotional response cannot be generalized across objects for different people. Rather, different people will determine the success of an emotion on a case-by-case basis. If that is so, then in what sense is emotional truth really *truth*, unless it can be conceived of as relative to persons?

Consider an example in relation to fictions. In chapter four I introduced Carroll's notion of *criterial prefocusing*, the notion that our emotional responses towards fictions are guided by

features of the narrative. Maybe prefocusing is enough to establish true emotional responses.

Is there a way in which my emotion toward a character could be false, by unsuccessfully characterizing its object? I have never really liked the character Brutus in Shakespeare's *Julius Caesar*. I was horrified that he would betray his friend, even though his friend was an egotistical despot. I remain unconvinced even at the end of the play when Marc Antony praises Brutus's honor and love of Rome. In other cases, one might sheepishly admit to admiring a wholly unsympathetic fictional villain. I am perversely intrigued by Nurse Ratched from *One Flew over the Cuckoo's Nest* (both the novel and the film) even though I *know* that you are not really supposed to like her. She is never presented in a sympathetic light, even when the patients on the psyche ward debate the merits of raping her. Perhaps due to my horror at this and other allusions to sexual violence, along with my empathetic feelings towards the nurse as the sole woman in a position of power in a male-dominated arena, I feel some compassion for Nurse Ratched. And yet, this seems like the exact opposite of how the narrative is intended to guide my feelings.

Prefocusing presupposes that there is some emotional guide to our experience of a fiction. An author, playwright, or movie production team intends for their audience to respond positively to certain characters and negatively towards others. This is what makes the fiction *work*. It seems like something has gone wrong for me emotionally if I watch *The Dark Knight* and do not feel sympathy for Batman and repulsed by the Joker. The information we are given about a character, the perspective and point of view of the narrative, the symbolism, language, lighting, music—basically every controllable aspect of a fiction—all work together to guide the audience to a certain response. Perhaps if we do not respond in the way the authors intended, then our emotion is false. In De Sousa's words, my emotion towards the protagonist or antagonist of the fiction would not be successful.

I think that prefocusing often does play an important role in shaping an audience's response to a fiction. However, I am skeptical that it can ground a theory of emotional *truth*—nor, indeed, does Carroll intend for it to do so. It is quite possible that a reader is sensitive to all of the narrative and linguistic ploys that an author incorporates into his or her work. No matter how clever and playful Nabokov portrays Humbert Humbert or how protective and caring Dexter Morgan becomes throughout his show, some audience members will simply never like these characters. This relates to a point I raised earlier with de Sousa's theory of correctness: each individual audience member approaches the fiction with his or her own values, beliefs, and personal/social history. These background features shape how we emotionally respond to the fiction. My strong conviction that a true friend should never betray their companion bars me from admiring Brutus's honor. This is so even as I recognize that Brutus was trying to do what was best for Rome, he acted honorably in the face of death, etc.

I remain unconvinced that there is a way to define emotional truth for fictions. Furthermore, it may be questionable that real-life emotions can have universal truth conditions, especially since our commonplace emotional responses lack *any* kind of prefocusing. Perhaps some other theory can help us to define emotional truth. However, any such theory runs the risk of overgeneralizing emotional responses and ignoring the diversity of potential responses that people can take.

I think that emotional *fit* can take the place of truth for explaining the rationality of our emotional responses towards fictions. The appropriateness of our emotions can be explained in terms of whether or not we have reason to respond in a certain way. Emotional fittingness may vary from person to person and context to context. This is important if we think of emotions as appraisals; not everyone will evaluate a particular person or situation in the same way, and so they

will not necessarily emotionally respond to it in the same way.

An emotion's fit can also be evaluated in terms of its *shape* and *size* (D'Arms & Jacobson 2000b). Shape refers to how an emotion represents the properties of its object. This can be a source of error in an emotional appraisal, when one's emotion presents an object as having (or lacking) a property that is does not (or does) actually have. An emotion's *size* refers to its intensity. I can be either annoyed or irate that my friend broke her promise to me, slightly downtrodden or overwhelmed by grief upon the death of my favorite movie star. We can also make errors in terms of an emotion's size, such as when the strength of an emotional response is out of proportion with its object.

Emotional fit helps us to understand when it is reasonable to have an emotion, but makes no claims concerning when we *ought* to have one. So fitness avoids the subjectivity problem that bothers theories of emotional truth. Whether or not our real-life fear is fitting will depend on context, our beliefs about the object, social norms, and its relation to us. Note that this is a different notion than the fittingness of *displaying* certain emotions, which may also be subject to social norms, both for real life events and in response to fictions (see Prinz 2004a, chapter 6). It might be strange to display intense anger when we are watching a film with relatives or people with whom we are not very close. But if we are having fun with friends and trying to make a point, then our angry display might be just fine. In some situations, it may be perfectly appropriate for us to respond angrily to a perceived slight in the real world, such as when we are with close friends. In other situations, the emotional display might not be fitting and expressing it would be extremely inappropriate.

Emotional fit can also explain my unusual emotional responses to Brutus from *Julius Cesar* and Nurse Ratched from *One Flew Over the Cuckoo's Nest*. I had decent justifications for my

feelings about Brutus Nurse Ratched. In that sense, my emotional responses were rational. My emotions fit their object. Part of the rationality of my emotions is due to my own personal background beliefs, values, and assumptions that I bring to my experience of the fictions. Not everyone will share my positions and so not everyone will have the same emotional responses that I do. These responses are *rational*—that is, reasonable—even if they are abnormal.

4.4. Emotional propriety

Virtually any scenario, person, or thing—real or fictional—can serve as the object of our emotions. As we saw in the previous subsections, the context in which we find ourselves will make those emotions fitting. Our understanding of emotional rationality goes beyond this distinction, however. Not only do we wish to know if it is appropriate or understandable that we experience an emotion, we also may want to know which situations we *should* or *should not* have particular emotional responses, as well as whether we can be held praise or blameworthy for doing so. I call this *emotional propriety*: the conditions under which we should be held morally accountable for our emotions towards fictions or real-life objects.

Consider the following example. A recent episode of *Game of Thrones* ("The Rains of Castamere") presents a gruesome and heart-wrenching scene in which several major characters—the heroes and moral compasses of the series—are brutally betrayed and murdered, marking the end of an honorable family's just rebellion. Distraught fans displayed their dismay online and on television immediately after the episode aired. They were outraged by the "Red Wedding" and what it implied for the rest of the series. These fans felt shocked, sad, angry, and betrayed, just as if someone they knew in real-life had met the Stark family's fate. In an interview after the episode

released, Martin commented that the scene was "like murdering two of [my] children. I try to make the readers feel they've lived the events of the book. Just as you grieve if a friend is killed, you *should* grieve if a fictional character is killed. You *should* care. If somebody dies and you just go get more popcorn, it's a superficial experience isn't it?" (Hibberd 2013).

Fans of the television show certainly did feel very strongly about the Red Wedding. The question is whether they *should* have. Martin clearly indicates that we should feel powerful emotional responses towards fictions. Otherwise our engagements with fiction constitute "a superficial experience." Indeed, we might find it odd and slightly off-putting to learn that a friend of ours finds a story interesting and enjoyable, but nevertheless remains totally disconnected from the characters. On the other hand, one might think that those of us who react very strongly to fictions are wasting "emotional energy" that might be put to better use towards things in the real world that deserve our attention, such as social injustices, real-life friends, natural disasters, etc.: the *real* sufferings of *real* people.

On what basis can we justify or condemn such emotions? One way to understand this is in terms of the *responsibility* for our emotions. It is often understood that we can be held morally praise or blameworthy only when we are responsible for our actions and moral judgments. Perhaps the same idea can apply to our emotions.

It might seem strange to hold someone responsible for their emotions. After all, it often feels like emotions are uncontrollable, and control may be a minimal requirement for responsibility. Robert Solomon argued that we *can* hold people responsible for their emotions (Solomon 1993 & 2004; see also Ben Ze-ev 2001). He notes that emotions are judgments; they do not happen *to* us. They are something that we *do*. Certain aspects of emotions are clearly involuntary. Alas, we cannot help blushing from embarrassment or experiencing the pangs of

grief. But other aspects of an emotion *are* voluntary. We can control the situations in which we put ourselves, the knowledge and information that we seek out, and our behaviors towards others. All of these influence our emotions. We can control the actions we perform on the basis of emotions (to a certain extent), like yelling when we are angry or wallowing in self-pity. It seems, then, that we can be held responsible for some aspects of our emotion. Note that according to the multi-level theory, emotions first arise as the result of an affective appraisal. All that I have said to this point suggests that this preliminary appraisal is not under our conscious, cognitive control. However, one could argue that the second, modifying aspect of our emotional *may* be under our control. In fact, there is a great deal of empirical and theoretical work on emotional regulation that suggests that our emotions can be modified, influenced, and controlled by cognitive feedback (see Gross 2007). If that is the case, then maybe we can be held responsible for them in some cases.

I will make one (perhaps obvious) suggestion: we should hold people responsible for their emotions when they *cause* (or are caused by) morally praiseworthy or blameworthy behaviors. Of course, what counts as praise or blameworthy will vary depending on the situation. Typically, though, I think that our responses can be evaluated in terms of the actual or potential harm or happiness they may cause. We can relate this back to the notion of fit. We can be held responsible for unfitting emotions as well as for fitting ones. Others may find our misshaped emotions blameworthy if they distort a situation inaccurately. We can also be held praise or blameworthy for an emotion's size: if a friend's anger is way out of proportion with its object (say, snapping at another friend for accidentally spilling coffee on her book), then we may blame our friend for her inaccurate emotion. Finally, the very reasons that we have for a particular emotion may be subject to praise or blame. We do not take accidentally spilled coffee on a book as a good reason for anger (especially extreme anger). We would say that this is not a good reason for a particular emotion;

the reason does not justify the outburst.

If I'm right, then there are situations in which we can hold people responsible for their everyday emotions. What about our emotions towards fictions? They seem to be a particularly interesting case, because their object does not actually exist in our world. Perhaps this is an ontological difference that *makes a difference* for emotional responsibility.

Actually, I don't think that it does. I think that we should hold people responsible for their emotions towards fictions in the same types of situations and for the same types of reasons that we would hold them responsible in the actual world. Consider a young woman whose sorrow over the death of a fictional heartthrob causes her to mope, lose sleep, or do poorly on a school exam. I think that many people would contend that she lacks control of her emotional responses and should be held responsible for the resulting actions. We would likely say the same thing if the woman's sorrow was the result of a real person's death, someone like a movie star or athlete with whom she has no personal relationship. The fact that the object of her emotion isn't a concrete particular does not seem to make much of a difference in this case. We will hold people responsible when their emotions lead to bad results or themselves or others.

We can finally return to Red Wedding example. Are our strong emotions appropriate in this case? First, we must ask whether the emotions fit their object. If we grant that the death of a beloved character may be an acceptable reason to feel sorrow, then the emotion is fitting and has a proper shape. One could argue, however, that it is mis-sized; the emotion may be far too strong for what the case warrants. To determine whether this is true, I think that we need to turn to our notion of emotional responsibility. Are we so caught up in the story and so distraught that we ignore other pressing duties? If so, then perhaps we do not *feel* as we should. Otherwise, I am inclined to say that the emotional response is appropriate.

5. Summary

I have argued in this chapter that we have genuine, rational emotional responses towards fictions. There were three distinct claims. First, our emotional responses towards fictions are genuine mental states. My arguments here involved dissolving the paradox of fiction. I argued that the functionalist assumptions that are implicit in the paradox are unfounded because we can be motivated to act in response to fictions. I also argued that emotions towards fictions play the same type of inferential role as real-life emotions, that our emotional feelings are genuine and non-illusory, and that we make genuine emotional evaluations of fictional objects. These claims paved the way for the SAV, but left open the question of the nature of our emotional responses to fictions.

My second major claim in this chapter is that our emotions in general should be thought of in terms of a multi-level appraisal of emotionally relevant objects. Our emotional responses involve low level perceptual and affective processes and slower cognitive processes that work together in a feedback loop. These processes combined make up a temporally extended emotional process, including conscious qualitative feelings. Because I do not think that our emotions require a belief in the existence of an object, we can say that our emotional responses towards fictions are genuine emotions.

Finally, I have argued that our emotions towards fictions can be rational. It makes sense that we should have emotional responses towards objects that we know do not exist. Indeed, our emotional interactions with fictions are often epistemically fruitful. Our emotional responses towards fictions are often appropriate—or, in my terms, fitting—according to an individual's personal experiences, background, beliefs about the fiction, as well as the content of the fiction

itself. Our emotional responses about fictions can be sometimes be proper or improper. This means that we can be held responsible for them in some cases.

The overall purpose of this chapter was to understand the nature and expression of our emotional responses towards fictions, especially those emotions that may be involved in our moral evaluations of the fiction. This was important to my project for several reasons. First, I suspect that emotions play a role in our moral judgments, an idea that I will further develop in the following chapter. Second, we saw in the previous section how propositional attitudes like beliefs can be understood to be genuine. This chapter solidified the SAV in terms of emotions as well as dissolved the elusive paradox of fiction. Finally, explaining my views on genuine, rational, fictional emotions will prove to be quite useful in chapters 7, 8, and 9 when I explain the three remaining puzzles of fiction: the sympathy for the devil phenomenon, the puzzle of imaginative resistance, and the question of moral learning.

Appendix

I would like to consider four potential objections to my multi-level appraisal theory of emotions. Several of these objections present challenges to any theory of emotions (2, 4) and some to appraisal theories in general (1, 3). Luckily, my theory can handle each of them.

Objection 1: Appraisal theories cannot explain emotional responses to novel stimuli.

All appraisal theories hold that emotions are responses to stimuli in one's internal or external environment that bear on one's wellbeing, or the wellbeing of a person or object on cares about.

The appraisal is typically understood in terms of a response to an emotion-proper property; we appraise stimuli with respect to the associations we make between them and past experiences that we have had with similar objects, people, or situations.

But what about entirely novel stimuli, objects that are not yet associated with a particular emotionally relevant property? How do we respond emotionally to *them*?

Novel stimuli would be a problem for *any* theory of emotions that has no argument for how we acquire knowledge and concepts of emotionally relevant properties. In §3.2, I suggested that some emotional information may be innate (as in the case of looming objects or darkness; see Damasio 1994, LeDoux 2012) and some may be learned. At least some of our emotional associations are likely acquired through social learning and experience. It has been the task of child development researchers to discover how this learning process takes place. The psychologists Dare Baldwin and Louis Moses (1994) studied how very young children gain emotional understanding from their parents (see also Stein 1996). Between the ages of 2 to 3 months, children are able to discriminate between happy and sad facial expressions. At the same age, they are able to track another person's line of vision by noticing that a parent is looking away from them. They can also follow pointing gestures to objects that are close by. These are three examples of early communicative capacities. Put together, they lead to a *social referencing* phenomenon which develops at around 8 to 12 months: when a child is presented with an unfamiliar object or person, she will typically glance toward a parent and then behave toward that object or person in terms of the affective cues that the parent displays. Based on these observations, Baldwin and Moses conclude that infants often spontaneously seek emotional information from a parent in order to help them know how to interact with the new object. This involves an implicit recognition that emotions have intentional and referential qualities (Baldwin & Moses 1994).

So we may not be able to respond emotionally to novel stimuli until we learn more about object and how our social group typically responds to them. Or, in other situations, if we learn to associate objects with certain positive or negative qualities through our own individual experiences. This is a question that bears on all theories of emotions, not just mine.

Objection 2: Emotions vs. moods vs. startle responses vs.…

One of the challenges facing emotion theorists is to define the difference between emotions and other sorts of affective states. What is the difference between full-fledged emotions and moods? Are startle responses full-fledged emotions? I have not presented answers to these questions in this chapter. However, I think that my process approach may help us to sketch a response to both.

Startle responses are sometimes not taken to be genuine or "full-fledged" emotions because they are short-lived and cognitively impenetrable (Carroll 2008). On my view, a startle response would be a quick, automatic appraisal of an object that results from low level processing (LeDoux's "low road") and so *does* count as a real emotion. A loud noise may trigger a sub-cognitive appraisal of something potentially dangerous, for example. The response may initially be processed by the same neural structures (the sensory thalamus and amygdala, in this case) but without the slower, cortical processing. Cognitive processing happens slightly later. This is when we will consciously experience fear—perhaps only for a few seconds, before we realize that it was just a car back-firing that made the loud noise, and not a gun. Perhaps the same sort of idea can apply to other low level affective responses, such as mimicry, mirroring, and emotional contagion.

Moods are a bit more challenging to explain. Moods are sometimes thought to be longer-lasting than emotions, free-floating and lacking an intentional object—in my case, not associated

with a core relational theme or emotion-proper property, or dispositional states as opposed to occurrent ones (see Prinz 2004a for an overview of positions).

There are problems with each of these views. Sometimes emotions can last a long time and moods can be short lived. They can also be dispositional; I may be disposed to undergo an angry state whenever I witness my neighbor going through my mail or an affectionate state whenever I see a French bulldog. The intentionality proposal would be the most difficult for the multi-level appraisal theory: how can we make affective appraisals if there is no *thing* to appraise? Yet it is difficult to explain how moods can be caused if they lack intentionality. Surely there has to be something that triggers a mood. Carroll (2003) proposes that moods may "spillover" from emotions so that we experience a state that is similar to an emotion, but applies to many different objects as opposed to just one (see also Prinz 2004a). This would help to explain how moods are caused and also how they come to be associated with certain feelings and appraisals; for example, we associate a cranky mood with something annoying or offensive and a cheerful mood from a random act of kindness or sunny day.

We can explain this in terms of my multi-level appraisal theory. Recall that, on this view, emotions are temporally extended. They involve different non-cognitive and cognitive appraisal processes in a feedback loop. We have an emotion when an object is initially appraised both via cognitive and non-cognitive processes. This may generate physical and physiological responses, including feelings. Some of these responses/feelings may be long-lasting and continue to influence our mental states throughout the day. Thus, if our original emotion was positive-valenced, we may continue to take in new information and process it in light of that earlier positive state. The result would be continued positive affect, but without a specific object: a good mood.

Objection 3: What about drug-induced emotions?

Imagine that there is a drug that you can take that will make you consciously experience the exact same phenomenological feelings you would if you were elated, ashamed, or enraged. Would these feelings be genuine emotions? It does not seem like they could they be if an appraisal theory is correct—after all, there is no appraisal occurring in this case. And yet many people will want to claim that drug-induced states are genuine emotions.

It looks like my any appraisal theory will just have to bite the bullet on this one and deny that drug-induced feelings are actual emotions. However, I don't see this as a problem for my view. I think that part of the appeal of cases like the drug induced affective state stems from our implicit commitment to a feeling theory of emotions. We generally have access to our emotional state through our emotions through their conscious feel: the positive, good feeling of joy, the uplifting feeling of pride, and down-trodden feeling of sorrow, or the frenzied feeling of anger. These feelings seem like the most important part of our emotion because they are the most salient. However, there are at least some reasons why we can question whether feelings alone constitute an emotion. We do not individuate emotions by their feelings alone. Rather, we tend to think that emotions play certain functional and inferential roles as well.

Imagine that you take a drug that makes you feel just like you would if you suffered some irreconcilable loss—but, in fact, no sorrowful event has taken place. That means we have an affective feeling without an appraisal, and so no genuine emotion. Does this go against our commonsense notion of what it means to experience an emotion? I doubt it. If someone were to ask you if anything is the matter, you may respond "I feel sad." The concerned friend would likely follow up by asking *why* you are sad. She would ask you to make some appraisal of an event or

object. You would likely respond by saying "Oh, I'm not really sad, I just took this pill that's made me loopy. I'll be better soon."

Compare this to an actual case of sorrow. When your concerned friend asks if anything is the matter, you say "I am sad" and then, perhaps, go on to explain the object or cause of your sorrow: "My best friend just moved to Prague and I'm afraid that I won't see her again for a very long time." The point is that genuine emotions may play certain functional and inferential roles that are explained by some object or cause of our feeling, not merely by the feeling alone.

Objection 4: How to explain pure-instrumental music?

Imagine listening to Mozart's *Clarinet Concerto*. You are in a particularly reflective state of mind, so you decide to keep track how this piece makes you feel: you find the tempo of the first movement uplifting, the introduction of the clarinet a pleasant change, the somber, doleful tones of the second movement rather disheartening.

We describe music as having emotional qualities. We also think that we have emotional responses to purely instrumental, non-representational musical works. But why? What is the source of these emotions—if they are emotions at all? This is especially puzzling if we think that emotions are a kind of appraisal since there does not seem to be any*thing* to appraise.

Questions like these have forced philosophers and cognitive scientists to scramble to find a way to explain our emotional responses to music. No one denies that we do have affective responses towards music. The question is whether these responses are genuine emotional states (denied by both Davies 1994 & Kivy 1999), illusory states (Prinz 2004a), some other kind of

affective state, such as a mood (Carroll 2003), or an "ineffable feeling" (Robinson 2010; see also Renero & Tullmann, in prep).

There are actually three related questions here. The first concerns whether or not our affective responses towards music constitute a genuine emotion (call this the *genuine emotion problem*). Second, we need to know in what way music can be expressive of emotions. Call this the *foundational problem*. Finally, we need to be able to explain the intentionality of our affective responses towards music. If most theories of emotions take them to be *some kind* of evaluation of our environment, we need to understand the nature of that evaluation. Call this *the intentionality problem*.

These problems are interrelated: an answer to the foundational problem will explain the intentionality problem, and the answer to the intentionality problem will explain the genuine emotion problem. A full explanation of these issues would, sadly, require a chapter (or book) in itself. Here I would just like to mention some possible responses to the foundational problem, each of which would be compatible with my multi-level appraisal theory.

What is it that we affectively respond to in pure-instrumental music? We often respond affectively to formal features of a piece: timbre, dynamics, pitch, tone, instrumentality, etc. The difficulty here would be to explain how it is that these formal features of music are capable of causing emotions (the intentionality problem). There are several ways in which we can explain this (Robinson 2010).

First, it might be that we have *appreciative* emotional responses towards music (Kivy 1999). We respond to the beauty of a work or the craft involved in constructing it. These appreciative responses take the work as a whole as their intentional object, but do not need to explain how one aspect of music causes a particular affective response. Another possibility is that

we respond affectively to instrumental music because we *associate* the sounds of the music with other emotional situations or objects (see Prinz 2004a). We associate marches with patriotism and hymns with piety. Or, we might associate a particular sound with a particular object, like an animal or human voice. Finally, we might have an aesthetic response to understanding the *structure* of a music work. We might feel relieved, surprised, satisfied, or unsettled by the introduction of certain instruments, themes, or dynamics (Robinson 2010).

I think that all of these possibilities are compatible with my multi-level appraisal theory. Importantly, our responses towards formal qualities of a piece—whether through association, appreciative emotions, or structural qualities of a work—each involve a kind of appraisal of the work or some related associated object. That is enough for the emotional appraisal process to get off the ground. I think that it is also enough for us to explain why these are genuine emotions with intentional objects. Unfortunately, that is a story for another day.

Chapter 6: Moral Appraisals of Fictions

1. Zombies, trolleys, & footbridges

Judith Jarvis Thompson's "trolley problems" have become just as famous as the ethical theories they were intended to comment on (Thompson 1986). The first trolley problem runs as follows:

> A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save these people is to hit a switch that will turn onto a side track, where it will run over and kill one person instead of five. It is okay to turn the trolley in order to save five people at the expense of one? (Greene 2007).

Researchers have found that approximately 90% of participants who are presented with the trolley problem agree that it is morally acceptable to flip the switch, thereby killing the solitary person while saving five (Bucciarelli et al 2008).

Interestingly, the responses flip in the footbridge problem, a variation on the trolley problem:

> A runaway trolley threatens to kill five people, but this time you are standing next to a large stranger on a footbridge spanning the tracks, in between the oncoming trolley and the five people. The only way to save the five people is to push this stranger off the bridge and onto the tracks below. He will die as a result, but his body will stop the trolley from reaching the others. Is it okay to save the five people by pushing this stranger to his death? (Greene 2007).

In this case, only about 10% of participants choose to push the stranger (Bucciarelli et al 2008). Why are the responses so drastically different in these two cases?

The neuroscientist Joshua Greene and his colleagues have drawn several important conclusions about the nature of morality from these two thought experiments. Greene and social

psychologist Jonathan Haidt (2002) take the responses to these cases as evidence that emotional intuitions play a role in our moral judgments, how we make moral evaluations. In another analysis of these studies, Cushman, Young, and Greene (2010; see also Greene et al 2001 & Paxton et al 2011) take the divergence in responses to the two problems as evidence of a dual-process theory of morality in which emotions and conscious reasoning play separate roles in our moral judgments.

Greene (2007) further concludes that the participants' responses to the thought experiments reveal the rational basis of utilitarianism and the emotional basis of deontology—a surprising conclusion considering Kant's own emphasis on reason. Greene explains his conclusion as follows: we are able to rationally calculate the costs and benefits of saving five people over one in the impersonal trolley problem because we are removed from the victim (the utilitarian response). However, the up-close-and-personal footbridge problem evokes a strong negative emotional reaction about the thought of pushing the stranger onto the tracks. It is challenging to rationally overcome this reaction. Thus, respondents tend to make the deontological judgment that it is unacceptable to kill an innocent person, even if doing so results in the preservation of five innocent lives.

If we take these researchers at their word, then it seems that thought experiments can play a significant role in gleaning ethical intuitions from everyday people. These intuitions are paramount in providing counterexamples to normative ethical theories, support for others, and evidence for how we actually make moral judgments.

Consider one more thought experiment:

> It's the zombie apocalypse. You and your band of survivors stumble upon a secluded farm in the woods. To your relief, the farm is safe, quiet, and free from wandering "walkers." The only problem is that the owner of the famer has decided to keep every walker he has encountered precariously locked in the barn near were your group sleeps each night. Tension and fear mounts amongst your group. A few members of the group do not want to destroy the walkers; they were people once,

> after all, and your group are guests on the farm. Others think that the walkers are a threat and need to be eliminated for the sake of the group. You have to make a decision—and fast. Do you agree to shoot the walkers, thereby breaking your promise to the farmer?

Here we have another thought experiment in which one person must decide the fate of a group. We must decide whether we should kill a few in order to save the many, or let the few live (even if they are walkers) with the high risk that the group will be destroyed as a result. The additional complication is that killing the walkers violates a promise the group made to the farmer.

This is not, strictly speaking, a thought experiment. It is a *fiction* (although I will argue that some fictions are thought experiments in chapter 9). Fans of the television show *The Walking Dead* will recognize this vignette as the central dilemma in the episode "Pretty Much Dead Already." The protagonist, Rick Grimes, must make a decision concerning the fate of his ragged band of survivors. Like the trolley and footbridge cases, this episode tests our moral values. Some viewers may take the rational response in the beginning of the episode (kill the walkers and break a promise for the greater good). However, as Greene's analysis would predict, these same viewers may change their decision after they emotionally connect with the farmers. We learn that one of the zombies is the farmer's wife, another is his daughter, another is his son-in-law, and (most importantly for the viewers) one is the young girl Sophia, who was separated from their group days before. Some viewers may switch their decision from a rational response to an emotional one upon receiving more personal information about the walkers and the farmer. Emotional closeness in this case takes the place of the *physical* closeness implied in the trolley/footbridge problems.

I have presented the trolley, footbridge, and "walker in the barn" cases to highlight the similarities between our moral evaluations of traditional ethical thought experiments and the moral scenarios presented in fictions. Most thought experiments *just are* fictions; they do not describe

actual people or events. It is plausible, then, that the audiences' responses to this case would likely have the same sort of normative and metaethical implications as the trolley problems. In fact, most fictional narratives are probably more realistic than standard philosophical thought experiments because they capture the complexity and messiness of real world moral situations.

The fact that we may be able to learn about morality from fictional thought experiments is important implication of the SAV. If we reason, judge, and feel about fictional event and people the way we do towards real-life situations. My goal in this chapter is to integrate what we have learned from previous chapters concerning our perception, social cognition, and emotional responses towards fictions into a comprehensive moral psychology of fiction. In previous chapters, we have determined that the mental states that could potentially be involved in our moral judgments of fiction—our beliefs, thoughts, emotions, judgments, feelings, etc.—are all stereotypical mental states. All that we have left to do is show that our moral judgments of fiction are also genuine.

A moral psychology of fiction shouldn't just concern itself with establishing whether or not our moral judgments towards fiction are genuine mental states. It must also explain our *actual* moral experiences. So establishing the SAV of moral judgments will only take up a short part of this chapter. The rest will be devoted to explaining our moral judgments in general and those towards fictional characters in particular.

In §2, I will make the case for a SAV of moral judgments about fictions. Next, I will sketch a version of *multi-level sentimentalism*, the idea that both affective and reasoning processes are involved in our moral judgments, based on the multi-level appraisal theory of emotions I presented in the previous chapter. We will then be ready to address another puzzle of fiction: the *problem of moral motivation.* I then bring together the various components of our emotional and moral

responses to fictions that I have argued for in the past several chapters, presenting a unified account of the moral psychology of fiction. I will conclude this chapter by returning to our discussion of the semantic and normative implications of emotions in order to see how this bears on the rationality of our moral judgments about fictions.

2. Genuine moral judgments

In the previous five chapters, I have established the SAV for our beliefs, perception, emotions, and social cognitive capacities. Our mental engagements with fiction utilize the same types of psychological states that we use in our everyday experiences. Any differences in our emotions, beliefs, and motivations to act towards fictional objects can be explained in terms of the intentional content of those states. This is what I mean by taking the fictional stance; we acknowledge that the objects of our engagement are fictional. This affects how we mentally respond to them.

We can make the same claims about our moral judgments about fictional characters and events. I will argue that moral judgments are constituted by both emotional and rational processes, including moral beliefs and judgments. There is no reason why our moral judgments are non-stereotypical if emotions, beliefs, and judgments about fictional objects are all stereotypical states.

Still, a proponent of the DAV could contend that there is something unique about moral judgments. Why would we make moral judgments about something that doesn't exist? Furthermore, aren't moral judgments supposed to be motivating? If that's right, and we have genuine moral judgments about fictions, then shouldn't we be motivated to act on them? I will address these questions in what follows. I will begin by discussing the nature of moral judgments.

I contend that our moral judgments about fictional things are *genuine*. This is true on any theory concerning the constituents of our moral judgments.

It is important to understand the nature of moral judgments so that we can make sure that there is nothing unique about moral judgments that might render those about fictions non-stereotypical. Most of the controversy concerning our moral judgments centers around three questions.[17] First, where do moral beliefs and motivations come from? We need to know whether our moral values are learned or innate, culturally influenced, or deduced from rational principles or emotions. Second, how does moral judgment work? Here, we wish to determine which mental processes and mechanisms are involved in how we actually go about making moral judgments (Haidt and Bjorkland 2007). Finally, what is different about our moral judgments about fictions, if anything?

I will primarily focus on the second and third question in this chapter, although I will mention the first when relevant. To begin, I will discuss moral judgments as types of occurrent, internal states or processes. There are two general views concerning how we make moral judgments. The first is *rationalism*. Rationalists argue that moral judgments are based on conscious reasoning and deliberating. Rationalism has a long philosophical tradition. Plato believed that our souls should be governed by our reason and that emotions bar us from making good moral decisions (Plato 1985). *Contra* Greene, Kant argued that the Categorical Imperative was both a universal and logical means of deliberation that a rational agent could employ to determine whether

[17] One issue I do not address in this dissertation concerns the difference between moral and other norms. In particular, I do not describe the moral/conventional distinction, a notion introduced by Eliot Turiel and his colleagues (Turiel, Killen, & Helwig 1987; see also Nichols 2005. They argued that moral norms differ from conventional ones in that they are serious, authority independent, and considered wrong due to reasons of fairness and harm to others. Although the moral/conventional distinction has been heavily criticized by moral psychologists (Bacciarelli 2008, Prinz 2007), it remains one of the few systematic methods for separating moral from other norms.

an action was morally acceptable (1785/1959). Contemporary Kantians like Christine Korsgaard follow suit; ethical judgments are based on practical reason (Korsgaard 1986). Utilitarians like J.S. Mill (1861/2002) and Peter Singer (1995) hold that our reasoning capacities allow us to weigh the costs and benefits of a moral decision in order to maximize utility. Finally, the social psychologist Lawrence Kohlberg conducted a series of (now highly controversial) studies that seemed to indicate that children and young adults develop increasingly objective and universal moral principles and can justify their moral decisions through conscious deliberation (Kohlberg 1981).

The details of how we actually *can* and *should* make moral judgments vary between rationalists. However, they are all committed to the idea that moral judgments is necessarily based on reason. We can contrast rationalism to *sentimentalism*. Sentimentalists argue that emotions play a significant role in how we make moral judgments. Contemporary sentimentalists trace their roots to David Hume, who famously stated that "Morality…is more properly felt than judged of" (quoted in D'Arms and Jacobson 2000b).

There are several different approaches to how emotions are involved in our moral judgments. I'll mention three of them here (but see also Hauser 2006, Huebner et al 2008, Nichols 2005, 2005, & 2008, Roedder & Harman 2010, amongst others). The *social intuitionist model* developed by Jonathan Haidt is one influential sentimentalist theory (2001; see also Haidt and Bjorklund 2008). This view states that we arrive at moral judgments through triggered emotional intuitions as opposed to conscious reasoning processes, although post hoc rationalization may be employed when people are called upon to justify their moral judgments in social contexts. In contrast, Jesse Prinz's *constructive sentimentalism* (2006, 2007) holds that emotions are both necessary and sufficient for moral judgments; the mere presence of certain emotions can make a neutral judgment a moral one (see Wheatley & Haidt 2005) and those who possess emotional

defects—i.e. psychopaths—do not make actual moral judgments (Blair 1995, Nichols 2005, Prinz 2007). We have already encountered Joshua Greene's *dual-process model* (Greene 2002, Cushman et al 2010). According to Greene, moral judgments are a non-natural kind; they involve both emotional and rational processes. Emotions act as a kind of "alarm-bell" for moral transgressions and are associated with personal actions. As we have seen, emotional responses typically lead to deontological appraisals. Cognitive moral reasoning is generally based on impersonal evaluations and is associated with consequentialist responses.

Like the rationalists, each sentimentalist has his or her own proposal on how emotions relate to reasoning processes and how they bring about moral judgments and motivate action. Nevertheless, they are all committed to the idea that emotions play *some* role in our moral judgments and behaviors. In the following section, I will propose a type of multi-level sentimentalism that is similar to Greene's dual-process model.

My own view is based on my multi-level appraisal theory of emotions, as well as the distinctions I have made throughout this dissertation between perceptual and cognitive processing. While that the proponent of the SAV need not adopt the theory of moral judgments that I propose here, it is beneficial for three reasons. First, I think that is good empirical support for some form of sentimentalism, but also both empirical and theoretical reasons why cognitive reasoning processes are *also* a significant part of our moral deliberation. Second, multi-level moral appraisal theory squares nicely with my previous arguments concerning social cognition and emotions. Finally, my view has great power to explain the remaining puzzles of fiction as well as capture our actual moral experiences with fiction.

3. A multi-level appraisal theory of moral judgments

While reading a novel or watching a film, you might have a vague sense that a character is morally vicious or has done a bad thing. You might have fast, seemingly automatic negative reactions towards that character for what she has done. This captures one sense of our moral judgments about fictions: they are quick, non-reflective or deliberate, and often emotionally-laced. However, our moral judgments of fictions can also be quite complex. Watching a realistic social drama like *The Wire* may cause a viewer to consciously, deliberately, and thoughtfully weigh possible judgments before reaching a firm conclusion as to the moral worth of a character or action. Our moral judgments about fictions may generally be a mixture of the two types: both automatic, non-reflective and emotional as well as deliberate and thoughtful. My goal is to develop a theory of moral judgments about fictions that accounts for the variety of our moral reactions.

In this section I will lay out one potential theory that can accommodate the depth of our moral judgments about fictions. I call it a *multi-level appraisal theory* of moral judgments because it is so closely related to my multi-level appraisal theory of emotions. I will spend the next two sections discussing moral judgments in general before exploring the implications of my theory for fictions.

3.1. Multiple appraisals

Like some other sentimentalists, I argue that our moral judgments are the result of both affective and rational processes. Moral appraisals can be basic, automatic, and unconscious, just like our

emotional appraisals. They may also be complex, deliberate, and conscious. Recall that my multi-level appraisal theory of emotions posits at least two separate appraisals that work together in a feedback loop. There was an initial affective appraisal of a stimulus in one's external environment or thoughts. At the same time, slower cognitive reasoning processes evaluate the stimulus through conscious judgments and inference drawing, the implementation of context sensitive information, background knowledge, and personal traits that promote, enhance, regulate, or modify the original affective appraisal. Cognitive processing is sometimes required in order to identify a state (this state is indignation as opposed to rage; guilt as opposed to shame) or the cognitive processing influences the initial affective appraisal so as to become an entirely different state.

I think that the same sort of process occurs with moral appraisals. Following other sentimentalists, I contend that we undergo an automatic affective appraisal of a stimulus, which is quickly followed by another cognitive appraisal and other cognitive processes, often including conscious reasoning. I hesitate to call the affective appraisal an emotional *intuition*, as Haidt does. First of all, I do not think that all of our affective appraisals are innate or genetically determined. As I suggested in the previous chapter, we can *learn* to respond to certain objects with emotions. Furthermore, 'intuition' is used by various authors in a myriad of ways. Haidt thinks that intuitions are unconscious emotional reactions (2002; Haidt and Bjurkland 2008). Walter Scott-Armstrong argues that intuitions are unconscious beliefs (2008). Cushman et al (2006) describe them as unconscious principles. Johnson-Laird and his colleagues define intuition as reasoning from unconscious premises (Bacciarelli et al 2008). Because of these various meanings, the term 'intuition' has the potential to engender obfuscation in the moral judgment debate. It is possible that theorists are talking past each other—e.g. Scott-Armstrong and Johnson-Laird, who argue that reasoning can be unconscious and Haidt, who thinks that reasoning is necessarily *conscious*. I will

return to this point momentarily.

I have argued that emotions are intentional; they are *about* something. Specifically, emotions are about an emotionally relevant object or property in one's environment. The objects that trigger an emotion are fixed independently of the response. An emotion is not solely individuated in terms of its characteristic phenomenology or expression, but rather by an appraisal of an object or scenario as having certain characteristics. For example, we feel guilty when we have performed some blameworthy action that causes others to feel contempt towards us. The affective appraisal of our action is both a reaction to external features of our environment and our response-dependent interpretation of it.

An initial moral appraisal is very similar to an initial emotional appraisal. Moral judgments may involve automatic affective responses to a stimulus after we think about or perceive it. This is similar to Greene's characterization of the "alarm bell" emotional response; we may have an automatic appraisal of a disgusting action as "bad!", "dislike!" or "avoid!" or an appraisal of a kind action as "good!", "like!" or "approach!" Some moral judgments may just consist of the initial emotional response, like when we passively watch a film and are too absorbed in the story to engage in further thought about a disgusting action (at least consciously). But further cognitive processing is often involved in our moral judgments. It is here that the differences between emotional and moral appraisals might appear.

Here's an example highlighting the difference between typical anger and morally significant anger. Suppose that your best friend dropped your laptop, shattering its screen and rendering it unusable. This makes you angry. Now suppose that you witness your friend drop your laptop as a way to get back at you for forgetting a date that you had set up ages ago. This *also* makes you angry.

Even though you experience anger in both cases, it seems to me that the first case is not an instance of moral anger. Rather, you just experience everyday anger at your friend's carelessness. You probably wouldn't say that your friend has deeply and unjustly offended you. The second case does seem to be a case of moral anger. Your friend has acted in a way that seems morally reprehensible. What's the difference, in terms of emotional and rational processing? The initial affective processing will be roughly the same in both cases; you perceive some event in which someone has caused you harm. This triggers an automatic negative appraisal. Other information makes the latter a case of moral anger—for instance, background information about the responsible agent and the context of the action. In the second case, we know that our friend intentionally acted offensively as a means of retaliation in the second case. We recognize that our friend has violated some moral norm, such as "do not intentionally harm another's possessions." The same may not be said about the previous example; your friend has not violated a moral norm and your knowledge of her action and intentions does not indicate that she has acted in a morally reprehensible way, even if she *did* make you angry!

I think that moral situations will likely be a matter of interpretation that vary on a case by case basis (see also Bachiarelli 2008). These examples show that the initial, automatic responses to value-laden stimuli in both moral and non-moral situations may be very similar. However, there are differences in later stages of cognitive processing of emotional appraisals and moral appraisals. The social context in which each arise may also be quite different. Both emotional and moral appraisals may involve an interpretation of the implications of an object or event. This draws upon our knowledge and beliefs about the case at hand.

3.2. Moral reasoning

Suppose that you are watching a superhero film. You experience a range of negative reactions towards the film's evil villain, including automatic emotional response. You may also make moral judgments about the villain's actions. These may be automatic emotional appraisals, as we saw in the previous subsection. These appraisals may never result in later cognitive reasoning. Generally speaking, though, some cognitive reasoning will take place when we morally condemn the film villain. We interpret the moral significance of the agent or event; we understand an action or event as morally significant and assess that significance in terms of a judgment. Assessing the moral worth of an action includes reasoning processes in which we (consciously or unconsciously) apply moral principles and background knowledge to a stimulus or scenario, draw inferences from prior knowledge, and make decisions based on our understanding of a situation.

My multi-level appraisal theory places cognitive reasoning processes in a much more prominent position than, say, Haidt's social intuitionist model. Haidt argues that reasoning in moral judgments involves mere "post hoc rationalization." Haidt considers moral reasoning to only involve conscious inference-drawing processes. I think that this begs the question against the rationalists.

Haidt's main justification for the conclusion that reasoning processes are unnecessary for moral judgments arises from his work on *moral dumbfounding* (2001). Consider Haidt's infamous vignette of brother/sister incest:

> Julie and Mark are brother and sister. They are travelling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decided that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think

about that? Was it OK for them to make love? (*ibid*, 814).

Many participants quickly make a negative moral judgment of the siblings' action when they are first presented with this vignette. However, participants were unable to justify their judgments with reasons when they were asked to do so. Many concluded by simply repeating that the action was "just wrong," which, as Haidt notes, is not really a reason.

It seems like participants in this study were unable to consciously reason to a moral judgment. It is possible, though, that they reason *unconsciously*—a possibility that Haidt does not consider. Thus, Haidt concludes that the moral judgment against incest could not have been reason-based. He then concludes that the judgment *must* be caused by an emotional intuition. This inference is invalid. If unconscious reasoning is possible (and there is good reason to think that it is) then the subjects could have quickly and unconsciously reasoned from a basic moral principle as the vignette unfolded. However, when called upon to *justify* their judgment, the subjects may not have access to their unconscious reasoning, or even why they hold a particular moral principle to begin with. Do we have a good justification for thinking that incest is wrong in the vignette? Some people probably do not; they simply conclude that it *is* wrong. This does not mean that this principle cannot be involved as a step in moral reasoning (e.g. "this is a case of sibling incest. But I think that sibling incest is wrong. Therefore, the action is wrong."). The basic worry is that Haidt conflates reasoning in *judgment* with reasoning in *justification*, which are two very different processes.

A sentimentalist could object that it is unrealistic to think that we make moral judgments via deductive processes, even if it is done unconsciously. The deductive model just does not match our actual experiences of making moral judgments (Harman et al 2010). Deductive reasoning may play a small role in our reasoning processes, moral or otherwise. It could be that moral reasoning

is a process of belief *coherence* according to which we attempt to make our information about a moral situation cohere with our moral values and principles.

Gilbert Harman, Kelby Mason, and Walter Sinnott-Armstrong (2010) describe such a process in terms of a connectionist model of reflective equilibrium. They propose that internal moral reasoning proceeds through a process of "adjusting one's beliefs and plans, in the light of one's goals, in pursuit of…a reflective equilibrium" (*ibid*, 239). This view is inspired by coherentist models of justification, in which a new belief must cohere (be consistent) with the subject's standing beliefs (Quine 1951). If it does not cohere, then either the proposed belief or the standing beliefs must be altered until a "fit" is achieved.

The authors base their theory on the psychologist Paul Thagard's theory of *constraint satisfaction* (Thagard 2000; see also Baljinder and Thagard 2003 and Daniels 1979). Constraint-satisfaction models of decision making predict that an emerging decision will be accompanied by a general shift towards coherence across any aspects of a dispute that are relevant to the matter at hand (Simon et al 2001). We base our practical and moral decisions on the relevant information of a situation in which we find ourselves or that we observe. This could include our beliefs, goals, emotions, and desires. We make use of all of these background features in coming to the appropriate answer to important decision. Our beliefs about a particular case at hand are weighed in light of our other relevant beliefs, emotions, etc. The shift towards coherence is the result of taking these various aspects into account. Our conclusion is that which fits best with all the relevant factors of a case from our own perspective.

On this view, we can change even our most firmly held beliefs. Each new proposal or occurence must be weighed and considered against our other beliefs, goals, etc. if any inconsistency arises between the information we have about the situation and our standing beliefs.

Some of our beliefs might need to be rejected in light of new information. This includes *moral* beliefs. When faced with Haidt's sibling incest case, for example, it's possible that participants must realign what they believe about incest to account for the information given about the case (there is no possibility for pregnancy, neither sibling was hurt, etc.). In some cases, participants eschew their former belief that incest is morally wrong; others will stick to their original belief because their negative emotional responses and negative stance towards appraisal weigh heavily on their judgment. Their moral appraisal would be very difficult to change.

The constraint satisfaction model applies to both thought experiments and real-life situations. We try to fit our information about a situation with our moral beliefs, even while engaging with a thought experiment like Haidt's sibling incest case. The same would be true if we confronted with a real-life case of sibling incest (although there may be quite a bit more information to take into account). If the constraint satisfaction process works in cases like these, it seems likely that it would also apply to our interactions with fictions. In many cases, our experiences with complex fictional narratives will be more similar to real-life moral situations than thought experiments that are relayed via short vignettes with little background information about the people or environment involved.

So while logical inference may sometimes be involved in moral reasoning, it is likely only one of the many types of reasoning processes available to us when making a moral judgment. Other types of reasoning include speculation on various outcomes, regulations or modifications of the affective appraisals, analyses of hypotheses and evidence, and, as we have seen, shifts towards coherent beliefs. All of this is done against the background of our overall goals, desires, and emotions, each of which may influence our final moral judgment. I think that this flexible, complex model of moral reasoning fits best with my multi-level appraisal theory. The multi-level moral

appraisal should be understood in terms of a feedback loop in which both affective and cognitive processes may influence the other.

Here is an example of what I have in mind. In their research on the reasoning processes involved in moral judgments, Cushman and his colleagues discovered that there were some cases in which participants do, in fact, consciously reason to a solution in a moral dilemma. One scenario asked: "Is it permissible for Evan to pull a lever that drops a man off a footbridge and in front of a moving boxcar in order to cause the man to fall and be hit by the boxcar, thereby slowing it and saving five people ahead on the tracks?" (Cushman et al 2006, 1083). A majority of participants answered with a positive response to the question and were able to provide sufficient justifications for their judgments. They consciously acknowledge that norms against intended harm played a role in their decision.

How did the participants reach this conclusion? Perhaps something like the following occurs: a participant, Jane, reads the vignette, and automatically has affective responses to it based on her mental representations of the information provided. This may include unconscious concern or distress for the potential victims and anxiety about having to make a difficult decision. The same information is processed via slower cognitive processes. Jane's long term memory activates any of her relevant moral beliefs. She considers and weighs various outcomes in the dilemma. Jane finally decides that pulling the lever would be acceptable. According to the coherentist model, Jane makes this decision because it coheres better with her emotional responses and beliefs than the decision to not pull the lever. This whole process may take place unconsciously and result in a relatively quick response in which Jane tells the experimenter that she would pull the lever. Or, as Cushman and his colleagues found was often the case, Jane may think through the moral dilemma consciously, deliberately weighing evidence and perhaps recalling a general principle that claims

that one should try to maximize the number of lives saved whenever possible.

A multi-level appraisal theory of moral judgments can best capture the underlying complexity and also the seeming automaticity of our moral experiences. This theory holds that subjects will generally have a fast, affective response towards morally good and morally bad actions and events. These affective responses are a relatively simple kind of appraisal of something we perceive or think about. An initial affective appraisal may be a positive or negative response that require further reasoning before arriving at an appropriate moral conclusion. Reasoning processes occur concurrently with the affective appraisal, but take longer due to cognitive and neural constraints.

Our initial affective appraisals account for the automatic negative feeling we have towards immoral actions or, alternatively, the positive feeling we have towards good actions. Moral reasoning processes at the same time give shape to these automatic responses, lending them validity and normative weight. This means that both reason and emotions are involved in our moral judgments. This makes my view a suitable middle ground between rationalism and sentimentalism.

3.3. Potential objections

Sentimentalist theories like mine face several standard objections concerning the role of emotions in moral judgments, and metaethical issues surrounding the objectivity of moral judgments, the possibility of error, and moral disagreement (see D'Arms and Jacobson 2000a & 2000b). I will respond to the first objection here and return to the metaethical concerns in §6.

There is plenty of evidence from empirical moral psychology literature to support the idea that emotions play *some* role in our moral judgments. Neuroimaging research suggests that

emotions at least co-occur with moral judgments (Greene et al 2007, Greene and Haidt 2002) and may influence the scope and intensity of our moral evaluations (Schnall, Haidt, & Clore 2008). There is even some evidence that emotions are necessary for moral judgments (Blair 1995, Prinz 2006) as well as sufficient for them (Wheatley and Haidt 2005). However, as others have pointed out, neuroscientific tools currently lacks the precision to determine when emotions occur in the process of making a moral judgment, *which* neural mechanisms are involved in making them, and which psychological states corresponds to exactly which neural activation (Prinz 2006, Huebner et al 2008). If emotions cause a moral judgment, then we should see the initial emotion mechanism at work followed by cognitive processing. This is what Haidt and Greene's studies purport to show. If emotions are necessary and sufficient for emotions, then, in theory at least, we could see affective mechanisms activated without the mechanisms that are typically involved in reasoning processes and a difficulty in making moral judgments if affective mechanisms are impaired (see Blair 1997).

We cannot know for sure whether emotions influence, cause, or constitute a moral judgment until neuroimaging technology catches up to theories of moral judgment and we have a better understanding of neural processing. Luckily, this is not a problem for my view. Making a moral judgment is a process that generally involves an affective appraisal *and* cognitive reasoning. Determining when and how the judgment "happens" may be unimportant. We have emotional responses to both morally praiseworthy and morally blameworthy actions. We also use cognitive processes to evaluate a moral situation. Each process influences and shapes the other. Furthermore, how we determine our own discrete moral judgment may be a matter of interpretation, just like it may be for our particular emotions.

Another rationalist challenge to sentimentalism may prove more of a problem for my view.

I have suggested that the initial step in moral judgments involves affective appraisals. Other sentimentalist theories, including emotivism (see Ayer 1952), have argued that moral judgments involve occurrent emotional responses. One problem here is that moral judgments do not always *seem* to be accompanied by affective responses. We can make calculated, rational judgments based on our moral principles without ever feeling an emotion. Other sentimentalists have explained this by arguing that we are *disposed* to feel emotions in response to morally salient actions, but that does not mean that we will always have an occurrent emotional state in response to them (Nichols 2004, Prinz 2007).

Another available response to this challenge is that our affective appraisals aren't always conscious; we will not always consciously feel *anything* while witnessing moral situations. Perhaps all that is consciously available to us is the reasoning processes involved in weighing options and reaching a judgment. It is still possible, however, that we still have some kind of low level affective responses towards moral situations even if we have no conscious awareness of them. This explains our intuitive experience of moral judgments according to which we can coolly, rationally make moral judgments without feeling an emotion. The unconscious appraisal view can also account for the psychological and neurological data that seems to show that affective processing of some kind does occur during moral judgments. It is a great benefit of my theory that it can account for both of these features of our moral experiences.

On my view, moral judgments are constituted by both emotional and rational processes and states. Importantly, there is nothing about moral judgments that would entail that we cannot make them about fictional scenarios. Indeed, most of the studies that cognitive scientists use to test our moral judgments are fictional thought experiments. Any differences between our moral judgments about real-life and fictional objects can be explained in terms of the judgment's *content*

(what they are about). The knowledge that the object of our moral evaluation is fictional may play a role in how we interpret it. We do not need to posit a distinct attitude to explain our moral judgments of fictions.

4. The problem of moral motivation

There is another way for a supporter of the DAV could undermine the idea that we make stereotypical moral judgments about fictions. Moral judgments in real-life generally motivate behavior, but those concerning fictions do not. I call this *the problem of moral motivation*: we tend to think that our moral judgments are motivating; if we believe that a certain action is morally wrong, then we will not perform it and may take measures to prevent it from occurring. Yet we do not seem to be motivated to act on the basis of our moral judgments about fictions. If morals are necessarily (or even only typically) motivational, why do we rarely act upon our moral judgments of fictional characters and situations?

The obvious response is that we know that the object under consideration isn't real. This explains why we won't be motivated to act in response to our moral judgments about them. This claim is right in one way and wrong in another. I have argued that we are motivated to act towards fictions, but that many of those actions never come to fruition. This is true even if moral judgments are necessarily motivating. Alternatively, it could be that moral judgments do not necessarily motivate action. Either possibility is compatible with the SAV.

I have already laid the groundwork for these two positions in previous chapters. In chapter 1, I argued against the standard functionalist argument that our beliefs about fictions are essentially unmotivating. I dismissed similar arguments concerning our emotions in chapter 5. My central

claim was that our beliefs and emotions towards fictions may motivate action, but these actions generally do not reach fruition due to conflicting beliefs and background knowledge that we have about the fictional work.

Moral judgments are often action-guiding. If we make the moral judgment that we ought to donate our superfluous money to the needy, then, in general, we will be motivated to act on that judgment by donating our superfluous income. All parties in the moral motivation debate accept this much. Divisions arise with respect to whether or not moral judgments *must* be motivating. Motivational internalists like Michael Smith (1996, 2008) claim that moral judgments are necessarily motivating. Smith holds a form of *weak internalism*: "If an agent judges it right to φ in certain circumstances C, then she is motivated to φ in C, at least absent weakness of will and the like" (1996, 175). This position contrasts with *strong motivational internalism*, the view that moral judgments are necessarily motivating and always result in action. Without the action, there really is no motivation. Moral judgments are intrinsically motivating on both accounts; the motivation to act is already implicit in the judgment. *Motivational externalists*, on the other hand, argue that moral judgments simply supply a reason that would justify one's acting on the basis of that judgment (McDowell, 1978, Schafer-Landau, 2003). Moral judgments are not intrinsically motivating. They require the presence of some other state, such as a desire, in order generate action.

The challenge is to explain weak internalism's "weakness of will and the like." When we suffer from weakness of will, we know that we ought to behave in some morally praiseworthy way, but, for some reason, we cannot bring ourselves to do so. For example, say that you have many vegetarian friends and read a decent amount of applied ethical theory that has convinced you that it is morally wrong to eat meat. Indeed, you have no really good reason why you should eat meat at all, other than that this is how you've always eaten and you enjoy the way that it tastes.

You recognize that habit and taste do not justify a morally blameworthy action. But you cannot bring yourself to give up meat. This means that you suffer from weakness of will concerning your moral judgment that eating meat is morally wrong; that judgment should motivate your action, but it doesn't.

Weakness of will is not the only deterrent for morally motivated action. Motivational internalists claim that, *ceteris paribus,* our moral judgments motivate action. In contrast to strong internalists, weak internalists contend that we do not always go through with the motivated action. Maybe you don't suffer from weakness of will about vegetarianism, but other factors prevent your motivation from coming to fruition. For example, you live in a place where vegetarian cuisine is not readily available, you cannot afford to buy sustainable products, or the rest of your family does not share your beliefs. All of these reasons make vegetarianism highly impractical for you.

Smith argues that the cognitive reasoning involved in our moral judgments motivates behavior. We could also think that the emotions involved in our moral judgments motivate behavior, as sentimentalists do. One will be unmotivated to act if she lacks the required emotional state. Evidence for this view comes from research on psychopathy; the psychopath lacks empathy and compassion, besides being defective in other basic emotions (Blair 1997). On the motivational internalist reading, this means that psychopaths do not make sincere moral judgments. They use the same words as normally functioning agents, but do not mean the same things by them (see Prinz 2007). Thus, they are not motivated to act morally because they lack moral concepts. On the motivational externalist reading, the psychopath is comparable to Hume's "amoralist," who has knowledge of moral concepts and makes sincere moral judgments, but has no desire to act on their basis (Nichols 2005).

As we have seen, both motivational internalists *and* externalists agree that moral judgments

generally motivate action. Weak internalists hold that overriding mental states and desires could prevent subsequent action. We can construct the following picture with this point in mind. For any moral judgment, we may be consciously and/or unconsciously motivated to act in some way, and so we do. It is also possible that we are consciously and/or unconsciously motivated to act in some way and *fail* to do so, due to weakness of will or conflicting reasons, beliefs, desires, emotions, etc. Finally, we will not go through with an action if we lack any sort of motivation to perform it. Any of these options are possible in a moral situation. For example, when we are deciding whether or not to take up vegetarianism, we may morally judge that eating meat is morally blameworthy and either *act* on that motivation, be motivated but *fail* to act (weakness of will), or entirely *lack* motivation to act (if externalism is true).

Emotions are sometimes thought to be fundamentally motivating. So if emotions are involved in moral judgments, then it would be natural to think that they are responsible for our motivation to act in moral contexts. I agree that emotions are often motivating. But that does not necessarily mean that reasoning processes are not also involved in motivating behavior, especially on my construal of moral reasoning. Cognitive processing may regulate, modify, or completely eliminate a behavioral motivation that arises from an affective appraisal. Our beliefs, values, and knowledge weigh prominently in determining whether or not we act on a motivation (see Schroeder, Roskies, and Nichols 2010 for a similar view). Conflicting desires and goals may prevent us from acting on a motivation. So while our affective appraisals may motivate action, they will not necessarily bring that action about; some further reasoning processes be also be required. Furthermore, understanding moral motivation in terms of affective appraisals does not necessarily eliminate the possibility that a belief on its own may be motivating (see McDowell 1978). This may be the case when we dispassionately consider a moral scenario and our emotional

dispositions are not activated. I don't think that such cases occur very often, if at all. But my view does not rule out their possibility.

We can now apply this discussion to our moral judgments about fictions. In the first chapter, I distinguished between being motivated to act by fictions *qua* fiction and being motivated to act *based on* fictions. The same distinction applies to moral motivation. We can be motivated to act in response to events and characters *in the fiction* or as the result of what we take away from a fictional narrative. The former motivations generally arise unconsciously as the result of low level perceptual processing. We perceive representations of fictional entities and have beliefs about them in the world of the fiction. We also make moral judgments of fictional entities that are relevant to the fictional world. For example, you may negatively judge the trigger-happy detectives in the television show *The Wire*. We certainly affectively appraise the detectives' actions. Since affective responses are at least sometimes behaviorally motivating, we may also have some unconscious low level behavioral reactions to the fictional story. These actions do not manifest because they are blocked by contradictory judgments and beliefs to the effect that the affective appraisal's object does not actually exist.

The Wire example shows that we can be morally motivated to act towards a fiction *qua* fiction. We may also be morally motivated to act in the real world *on the basis of* our moral judgments about a fiction. Here is an example of what I have in mind. Say that you just saw the movie *Fruitvale Station*, which recounts the last day in the life of Oscar Grant III, who was killed by a Bay Area Rapid Transit (BART) police officer in 2009. The film captures the racial tension and discrimination that still plagues American society. The use of real cell phone footage of the murder adds to the immediacy and reality of Oscar's story. As A.O. Scott, the *New York Times* movie critic, points out:

> The climactic encounter with BART police officers erupts in a mood of vertiginous uncertainty, defusing facile or inflammatory judgments and bending the audience's reflexive emotional horror and moral outrage toward a necessary and difficult ethical inquiry. How could this have happened? How did we — meaning any one of us who might see faces like our own depicted on that screen — allow it? (Scott 2013).

Scott's point is especially relevant at the time the film was released—during the midst of George Zimmerman's racially-charged trial for the death of Trayvon Martin, another young African American man (*Fruitvale Station* was released on July 12th, 2013; Zimmerman was proclaimed innocent on July 13th).

A sensitive American viewer may be motivated in some way to act based on their judgment about Oscar's death. How would this moral judgment motivate action? On my view, a negative affective appraisal of the film is a component of the moral judgment. Negative emotions (sorrow, betrayal, loss, etc.) may motivate real world action. It may have been the filmmaker's goal to incite real world action through evoking moral emotions in an audience. Cognitive reasoning processes, including inference drawing from our beliefs about racial tension and profiling in America, may shape and influence to our affective appraisal. In some cases, a moral judgment about the film may incite a viewer to act. She may be motivated to take part in protests against the court's ruling in the Zimmerman trial, learn more about New York's stop-and-frisk program, or in other ways oppose laws that promote racial profiling. In other cases, one's sincere moral judgment concerning Oscar's story may be deterred by such factors as weakness of will—a viewer might know that racial profiling is wrong and that she should do something to stop it, but ask herself what she could possibly do in the face of such broad institutional discrimination. Or conflicting desires, beliefs, and other circumstances may prevent a viewer from acting on his or her judgment. Maybe the viewer was emotionally affected by Oscar's death onscreen, but for political reasons does not feel

obliged to act in any way.

Even though one's moral judgment of Oscar's story motivates us to act, it does not motivate us to act on our moral judgment of the fiction *qua* fiction. Instead, I think that our response towards the fiction may serve as the basis for a more general judgment along the lines of "racial profiling is morally wrong." Our morally motivated action is likely based on *this* judgment, even though it was brought about by the initial moral judgment of Oscar's story.

The *Fruitvale Station* example may seem cheap. It is, after all, based on a true story, so it is natural that it incites morally motivated action. However, I contend that the same picture of moral motivation applies to fictions that are not based on a true story, like our judgments against the elusive Willy and the racist neighborhood representative Karl Lindner in *A Raisin in the Sun* (another story of race relations in America), or even against a story that is entirely removed from our actual world, like *Avatar* or *The Lord of the Rings*. In these two cases, we might form a judgment that it is morally unacceptable to harm innocent, sentient life forms solely for self-gain that may motivate later action. Indeed, we often think that fictions illustrate important cultural and psychological concepts and ideas that we can apply to the real world. Fictional dramas such as *A Raisin in the Sun* and *The Wire* may help viewers gain new perspectives on real-life race relations, perspectives that may motivate action in their daily lives.

5. The moral experience of fiction

We are finally ready to piece together the various components of my moral psychology of fiction.

In chapter 2, I discussed how we perceive fictional objects. Chapter 3 presented the key notion of taking the fictional stance, according to which we recognize that the objects of our engagement are fictional (they are non-actual, they are created, they have particular dependence relations) while we simultaneously "see" them as the objects that they represent. Chapter 4 discussed how we understand the mental states of fictional objects. Chapter 5 delved into our emotional responses to fictions in more detail. This chapter has targeted our moral judgments and actions. The three latter chapters have heavily relied on multi-level psychological theories in order to best capture both the immediacy and complexity of our moral experience of fictions.

Let's consider an example of how this whole process works, using a scene from the film *District 9*. In this story, a space ship stalls over the city of Johannesburg sometime in the near future, stranding thousands of intelligent but malnourished aliens (called "prawns" for their crustacean-like appearance) in the city. The frightened humans round up the aliens into District 9, a slum kept apart from the rest of the city by high fences, guns and missiles. The protagonist, Wikus Van De Merwe, executes the order to evict the aliens from District 9 to the concentration camp District 10, further away from Johannesburg. As he and his team go about District 9 passing on the eviction notice, we witness the Wikus and others commit various acts of atrocity against the aliens, abusing and killing them indiscriminately, even "aborting" a nest of alien eggs.

Many viewers of *District 9*, recognizing the film's not-so-subtle commentary on South African race relations, may experience a range of moral emotions during this scene: disgust at the killings of the aliens, anger at the cruelty of the humans, and sympathy for the relatively helpless aliens. But even without the social association, we might feel a great deal of uneasiness and anger at the humans' actions. According to my multi-level appraisal theory of moral judgments, these emotions partly constitute our moral judgments about the characters in the film. We feel

indignation towards the humans' xenophobia, constituting our moral judgment that their behavior is wrong. We may also *reflect* on our indignation and moral belief that we should not deliberately harm other life forms without just cause, and especially not intelligent ones. We feel anger or disgust at the humans and draw conclusions about their actions based on our knowledge of this particular scene, what we know about the fictional world, and our own real moral values and principles.

Our moral judgment of the humans' actions occurs as the result of taking the fictional stance. We begin by perceiving the objects on the screen in a non-illusory way; we perceive representations of aliens, humans, eggs, shanties, etc. We know that the humans and prawns in the narrative are fictional. They do not actually exist in our world, they were created by the filmmakers, and they depend on the film, filmmakers, and viewers in order to persist. This knowledge forms the backdrop against which we take the fictional stance. We may not consciously acknowledge that the objects in *District 9* are fictional, but that status shapes how we think about and respond to them. At the same time, we use the 'is' of fictional transformation when we think about the fictional objects, "seeing" fictional entities as representations of objects.

Taking the fictional stance allows us to understand that the prawns and humans in the narrative are objects towards which we can have emotional responses. We also treat them as objects that possess mental states. My modified TT states that both perceptual and cognitive inference drawing are involved in how we understand fictional characters. We also may use social referencing and knowledge of a character's social context, personality, and background in order to inform our mental state attribution. Both perceptual/affective and more complex cognitive processes work together to shape our understanding of fictional entities and our own emotional responses to them.

The multi-level appraisal theory of emotions goes hand in hand with my modified TT. We make a quick, automatic affective appraisal of the humans in *District 9*. We also emotionally appraise the humans based on our beliefs and knowledge about their character and actions. Here, too, the automatic affective appraisal and slower cognitive appraisal work together to form a temporally extended, flexible emotional process that shapes how we feel about the characters. Our emotional responses about the humans also set the stage for our moral judgments of their conduct.

This is the standard process of our moral evaluations of fictions: we take the fictional stance, perceive fictional representations, attribute mental states to fictional characters, and have emotional responses to them. All of these processes shape our moral judgments of a character. All of the mental states involved in our engagements with fictions are genuine, stereotypical mental states. This means that my standard attitude approach is a viable contender to distinct attitude views.

6. Normative implications

I now want to return to the metaethical challenges to sentimentalism that we encountered in §3: the possibility of moral error, disagreement, and objectivity. I will draw upon the concepts of emotional rationality we discussed chapter 5 to respond to these challenges. Since moral judgments are constituted by emotions, it is important to understand how our emotions are used to evaluate the world.

Fitting emotions are those that we have some reason to feel; they obtain in virtue of a response-dependent property of an object. For example, we find an object in our environment

frightening because we judge that it may cause harm to me or someone I care about. However, just because one's emotion may *fit* its object does not mean that it is morally appropriate to experience it. Emotional truth and propriety capture the normative dimensions of our emotional responses. *True* emotions are successful or unsuccessful depending upon whether they accurately describe their object. For instance, feeling fear towards a harmless bunny would be unsuccessful, and so untrue. Finally, emotional *propriety* captures the moral dimensions of our emotional responses and how we can be held responsible for them. This was the case in instances in which our emotional responses were harmful to ourselves or others.

Sentimentalist theories hold that emotions figure in our moral judgments (D'Arms and Jacobson 2008). One challenge for these views is to explain *which* emotions are appropriate to a particular moral scenarios, but without implying that the emotion itself always has this normative implication. Another challenge lies in explaining how emotional responses can be objective. Emotions are subject-dependent; different people will respond with different emotional appraisals to particular objects according to their own knowledge, values, and beliefs. If *that's* true, then it may be difficult to see how there can be objective, world-based or universal emotional appraisals that people can disagree about and get wrong.

One worry is that if sentimentalism is true, then this subjectivity about our emotions carries over into our moral judgments. All sentimentalist theories adopt the "response dependency thesis" (RDT), which states: to think that *X* has some evaluation property φ is to think that it is appropriate to feel F in response to X" (D'Arms and Jacobson 2000b, 729). For example, projectivist theories such as Mackie's error theory (1977) and Simon Blackburn's quasi-realism (1985) hold that moral evaluations to not correspond to objective features of the people or situations, but rather are projections of their values onto the world. Allan Gibbard's (1990) theory of norm expressivism

identifies moral wrongness with the acceptance of a norm that takes anger as an appropriate response towards a particular action. I evaluate my friend's broken promise to me as morally wrong if I have internalized a norm that it is acceptable for me to be angry with her for breaking a promise to me.

D'Arms & Jacobson argue that sentimentalist theories like Gibbard's do not distinguish between the different ways in which our emotion may be appropriate for its object. They conflate an emotion's *fit* of and its *propriety*. But it seems like these concepts can be separated. Fear towards both spiders and battlefields may be fiting if it is intelligible that we should feel that way; one could sustain serious injury on a battlefield and we think that spiders are small and icky. Envy of a colleague's promotion may be fitting if you are up for the same promotion. Feeling amused by an off-color joke may be fitting if the joke is told in a charismatic way. Again, the emotion need not be an *appropriate* response in order to be fitting. An emotion can be fitting even if it is morally unjustifiable all things considered. That off-color joke that I find so amusing might be highly inappropriate in certain contexts, yet I may nevertheless have a reason to be amused by it.

This means that sentimentalist theories need some independent way of distinguishing between when we have a reason to respond emotionally to an object and when we *should* feel the emotion—i.e., when our emotional response constitutes a moral evaluation. To do this, an adequate sentimentalist theory must distinguish between and emotion's fit and its propriety: when it is reasonable or intelligible to experience an emotion and when it is morally appropriate to do so. Luckily, the notions of emotion proper-properties and core relational themes provide this account for my multi-level appraisal theory. These notions provide a non-circular way of explaining the response-dependence of the emotions involved in moral judgments (contrast with Wiggins 1987 and McDowell 1998; see also Prinz 2001 & Morreall 1986)). A scary object is one that causes fear

in us because we perceive it as a potential threat or source of harm. A disgusting object is one that causes a feeling of disgust in us because we perceive it as literally or metaphorically toxic, malodorous, unhealthy, or indigestible. A shameful act is one that causes me undergo shame because I believe that I have violated a moral norm or social taboo. This response depends both on the subject (is subjective) and on social norms. This makes it objective (at least culturally).

On my view the original affective response of a moral judgment is an appraisal of an object. This is not an endorsement of the object's moral *propriety*. It is an appraisal in terms of fit. When we appraise an object as being scary, we form a mental representation of it; part of our concept of this object might include properties like 'being dangerous,' which is then interpreted as scary. This will generally lead to avoidance behaviors and an unpleasant, scared feeling. But it will not necessarily be a judgment about whether or not we *should* take that stance towards the object. A later, cognitive evaluation will be brought to bear on the propriety of whether we should feel afraid of something and whether we can be held responsible for our responses to it.

Our moral judgments are subjective in the sense that different people will respond to and interpret emotional objects differently. But they are not completely subjective. We can make errors in our moral judgments and we can disagree with other people about how an morally relevant property applies to an object.

Recall D'Arms and Jacobson's contention that emotions have both a *shape* and *size* (2000a). An emotion's shape corresponds to how an emotion represents an object as possessing (or lacking) a property that is does not (or does) actually have. One source of moral error arises from misinterpreting the emotion's shape. We may misattribute properties to an object that it actually does not have. So, for, example, my envy of my friend's promotion might lead me to think that she does not deserve the promotion or that I am a better candidate than she is, when in fact

neither of these beliefs are true. My envy is misshaped because it fails to accurately describe my friend. An emotion's *size* refers to its intensity; I can be slightly annoyed that my friend broke her promise to me or I may be quite irate about it. I can be slightly downtrodden by the death of my favorite fictional character or I can be overwhelmed with grief. We can also make an emotional error in terms of our emotion's size. Consider our friend from the previous chapter, who became extremely involved in her favorite television shows. She was devastated when bad things happen to her favorite characters and elated by a villain's undoing. We may want to say that our friend has made an emotional error; her emotions are out of proportion with their object. Specifically, the object does not warrant the intensity of her emotion.

We can also go wrong in our moral reasoning. We can straightforwardly make errors in our thought-processes that lead to a moral evaluation or we can misinterpret important contextual information. Our emotions can also be in error if they influence us inappropriately, as in cases of bias, weakness of will, or self-deception (see Barnes 1997 & Mele 2001). These may all be instances in which our moral judgment does not cohere properly with our other standing beliefs, desires, and values.

Both the affective and cognitive appraisals involved in moral judgments may be sources of moral disagreement. People can disagree about how and to what object we apply an emotionally relevant property (see Nichols 2004 and 2008). I might think that a particular object or action is disgusting (like when people clip their nails in the subway car), whereas my friend might not find the action disgusting at all. We disagree about the emotion's shape. And even if we do both agree that trimming one's nails in a subway car is disgusting, we might disagree about the appropriate size of our disgust. I might be totally grossed out by the nail clippings, whereas my friend is only mildly put off by them. Finally, we can also disagree with respect to our background beliefs,

knowledge, and values concerning a moral situation. This would be the source of dispute between many people concerning current social and political issues like drone warfare, gun control, abortion, and the legalization of recreational drugs. We each come to the table with our own beliefs, knowledge, and emotional reactions to a particular moral scenario. These factors are all potential sources of moral debate because they correspond to objective features of the emotional object.

We can apply this discussion to our moral evaluations of fictional characters. Recall that there may be a sense in which fictional narratives call for a particular emotional response. Following Carroll, I called this *criterial prefocusing*. The emotional effect that the author, dramatist, or film production team works to achieve is brought about by prefocusing. The author manipulates the themes, language, symbols, etc. of a literary work, or music, lighting, dialogue, editing of a film or play. However, as we noted, not every person in an audience will experience the intended emotional response. I explained this in terms of the cognitive factors that often go hand in hand with an emotion: the beliefs, values, and thoughts about the moral situation or person that is presented in a fiction.

Our moral judgments of fictional characters can be pre-focused in the same way as emotions. If our emotions can be prefocused, then our moral judgments can also be shaped by the manipulation and influence of our emotions. For example, authors may shape our moral evaluations of protagonists by engaging our positive pro-social emotions: empathy, sympathy, compassion, identification, etc. We feel compassion for and sympathize and identify with the likeable characters in a fiction (see Carroll 2008, Plantinga 1999, Smith 1999 etc.). This is because the narrative presents these characters in a positive light, emphasizing their likeable, pitiable, or

admirable qualities so that we desire for them to succeed in their endeavors and worry, fear, and mourn them if they do not. Our positive emotions make us morally sympathetic to the protagonist.

A further interesting question concerns whether we can make inappropriate or unfitting moral judgments about fictions. Presumably our moral judgments about fictions can be in error in some way, if they can in real-life. There are three interesting types of case here. The first type of incorrect moral evaluations is not, strictly speaking, an instance of moral error *qua* fiction. Sometimes we sympathize with characters that we would normally abhor in real-life: Humbert Humbert, Dexter Morgan, Beatrix Kiddo, Satan. This is a perfectly normal response to the fictional character—possibly even an appropriate one. The artists intended for their audience to sympathize with these morally questionable characters. Does this constitute a moral error *all things considered*? Is it morally wrong to sympathize with fictional devils? I will return to this question in chapter 7 when I address the sympathy for the devil phenomenon. Another question concerns whether our moral evaluations are improper even if they are fitting. I will address this in chapter 8 when dealing with the puzzle of imaginative resistance.

A second type of case of moral error about fictions occurs when our moral judgments about a fictional object are unfitting. Moral judgments may not correspond to how the author portrays a character or event. Imagine that when I read the *Harry Potter* novels and, instead of admiring and cheering on Harry, Ron, and Hermione, I instead secretly admire the evil Lord Voldemort. Instead of morally condemning Voldemort, I quietly congratulate him in his renewed efforts to create a world of magical purity (I do not admit this to my friends, or, if I do, they think that I am just being intentionally, rhetorically perverse). It seems that I have made a moral judgment of the character that was unintended by the author.

Voldemort is supposed to be a highly unsympathetic character. Almost everything we know about him confirms this point: he kills innocent people—including children—he uses his followers, he is disturbingly racist, and he fits every criterion for anti-social personality disorder. There are very, very few instances in the story in which one might find oneself feeling sorry for Voldemort. There are some solitary examples of Voldemort's difficult upbringing. Still, it seems clear that we are supposed to be morally repulsed by the Dark Lord and his callous, violent followers.

In this case, moral error can occur in one of two ways: our emotional response may not fit the object—the emotion's size or shape may be wrong—or our cognitive evaluation of the object may be flawed. I think that the size of our emotional appraisal of Voldemort will not much matter in terms of moral blameworthiness. It is either wrong to sympathize with Voldemort or it isn't. It does not make much of a difference if one's sympathetic emotions are weak or strong. The emotion's shape *is* relevant, though. Imagine that one's sympathy for Voldemort's upbringing causes me to attribute other sympathetic features to him (like Voldemort's loneliness and sorrow after being abandoned by his father) or causes one to overlook other morally questionable qualities or actions.

We can also go wrong in our cognitive processing of the Voldemort's actions. Perhaps in judging Voldemort's "Magic is right" campaign (the magical equivalent of Thrasymachus' "Might is right" stance in Plato's *Republic*) the reader not fully consider the consequences of this position. Maybe she does not remember important details of what brought the campaign about. She may think that Voldemort's campaign is consistent with her other moral beliefs, when in fact she has not fully considered my other moral principles, such as protecting the rights of others and keeping innocents from harm. Or maybe she suffers from weakness of will and simply concludes that this

is the easiest path to take for universal happiness. In any case, others will likely conclude that the reader has made a mistake somewhere in her moral reasoning, making her a less-than-ideal moral agent.

We have made a moral error if we misunderstand a character's qualities and misinterpret their actions. Yet the "sympathy for Voldemort" example raises another interesting question in terms of the moral *propriety* of our judgments of fiction. Can we be held morally *responsible* for our judgments of fictional characters? Can I be blamed for admiring Voldemort? Let us suppose, unlike in the previous example, that a positive judgment of Voldemort does not arise due to some emotional or cognitive error. The reader's emotional response towards Voldemort is fitting in terms of both size and shape—it is not that she misattributes positive qualities to Voldemort and ignore others, or has a disproportionate positive response towards him. Rather, she simply thinks that Voldemort's actions are justifiable and that he has been misunderstood by the other characters and Harry Potter fans. In other words, the reader's moral values are incompatible with the negative portrayal of Voldemort and most people's negative evaluation of him. Given Voldemort's negative portrayal, I take it that many people would be uncomfortable with this evaluation, even if she never acted on it or in any way behaved like Voldemort or his followers. They would find this reader strange, morally perverse, unusual, and potentially dangerous.

I have suggested that there might actually be a great deal of variation in how audiences morally evaluate fictional characters. This is because each audience member brings her own beliefs, values, and desires to her engagement with the fictional narrative. This variation carries over into one's moral appraisal of the work; if we have different emotional reactions towards fictional characters, and emotions comprise moral judgments, then our moral judgments will vary

to the same extent that our emotional appraisals do. My view, then, seems to imply a moral relativism, similarly to some other sentimentalists (see Harman 1996, Nichols 2004, Prinz 2007).

However, my theory can also accommodate moral realism. On my view the moral judgment process continues beyond the affective appraisal. It involves cognitive processing and reasoning as well. These processes can factor in to our moral judgments and even modify the initial emotional responses. While our emotional appraisals are required for moral judgments to occur in the first place, there is no *prima facie* reason why we have to accept them as moral truth. Our reasoning can modify and guide our initial judgments. If that is the case, then moral realists can take heart in a theory that has a place for emotions in moral judgments. The reasoning processes might serve as the normative basis for one's moral values. This could lead us to judge the pro-Voldemort reader's positive moral judgments of this character as inappropriate.

I will to conclude this chapter by addressing a related point on moral propriety. In his forthcoming paper, "Pleasurably Regarding the Pain of Fictional Others," Aaron Smuts attempts to make the case the intrinsic wrongness of taking pleasure in an innocent character's suffering. It is possible for someone takes pleasure in the portrayal of Hermione's physical suffering when she is being tortured in *Harry Potter and the Deathly Hallows*. Smuts asks us to imagine that the reader takes a *secret* delight in Hermione's suffering. Suppose that the reader never tells anyone about her delight and her delight is never manifested in her behavior. With these constraints in place, how could it possibly matter whether a reader takes pleasure in Hermione's suffering?

Smut contends that those who think that there is nothing inherently wrong with taking pleasure in the suffering of fictional characters are committed to the *fictionality thesis*. The fictionality thesis holds that fictions are autonomous fantasies in which we are merely external spectators to the action (*ibid*, 5-6). Importantly, as Thomas Hurka points out, we are perfectly

capable of imagining that some immoral acts could potentially be acceptable in a fiction (Hurka 2001). Unlike in surrogate fantasies, which we hope to come true, the audience knows that the fictional actions are generally impossible in our world. This allows her to imagine all sorts of scenarios that she normally would not and consider moral stances that she might typically ignore or abhor.

It follows that taking pleasure in fictional suffering is "perfectly harmless" (*ibid,* 8). Smuts quotes Susan Feagin, who argues: "the freedom of imagining is freedom without responsibility. Pleasure in what one imagines can be as fickle or as base as one likes, without consequences" (Feagin 1984, 50; quoted in Smuts *in prep*, 8). Fictions *aren't real* and even if fictional objects exist in some way, they are not the sorts of things that can be harmed.

Despite the claims of the fictionality thesis, Smuts argues that taking pleasure in the suffering of fictional characters is inherently wrong. In order to understand this point, imagine a (potentially) real-life example:

> Imagine slipping on a banana peel at the local supermarket. You spin to brace your fall, but on the way down the corner of your mouth catches on the sharp lip of a shelf. It rips your mouth wide open. Through the gaping flesh of your torn cheek, most of your teeth are visible. You scream in agony. The blood fills your mouth, pours down your face, and pools on the floor. The paramedics arrive quickly. As they tend to your wound, a crowd gathers. Some softly snicker; others just watch. Unbeknownst to you, most of the crowd admires the scene, taking pleasure in your sobs of pain and the sight of the red blood oozing out of your wound (*ibid*, 10).

We are supposed to take it that no harm can come to the victim from the spectators' pleasure: they do not act on their mental attitude and the victim does not even know that they feel it.

On what basis are the spectators' actions harmful? Smuts suggests that appealing to a normative ethical theory can help us understand the wrongness of this scenario. The fictionality thesis assumes a consequentialist theory of moral wrongness: morally wrong actions are those that cause harm. Yet, in the grocery store example, there does not seem to be any harm caused by the

spectator's pleasure. One could argue that there is *some* harm done; maybe the observers who take pleasure in the victim's suffering harm *themselves* in some way—their callous pleasure makes them worse moral agents, what goes around comes around, etc. But, as Smut points out, we are not so much concerned about the harm done to the observers. We are concerned about the *victim*. It is the harm done to *him* that needs to be explained. The problem is that it's not clear that he does come to any harm as the result of the spectators' pleasure.

It could be that taking pleasure is morally wrong on other normative moral theories, such as deontology or virtue theory. For example, a Kantian might argue that is it always wrong to take pleasure in the pain of others, *period*. We are obligated to treat others with dignity and respect; taking pleasure in another's suffering undermines their worth as a rational agent. Alternatively, one could take a virtue theoretic stance according to which we should model ourselves after the behaviors of virtuous exemplars, in our attitudes as well as in our actions. Surely an ideal moral agent would not take pleasure in the suffering of others.

Perhaps the same is true for straightforwardly fictional examples. Smuts asks us to consider a case comparable to Moore's Two World example:

> Imagine two worlds, each having just one inhabitant, say a sole survivor of a nuclear holocaust. In world A, the survivor spends her free time thinking nice thoughts. She often imagines casts playing with rubber bands on sunny window sills. In world B, the survivor lives a similar life, but rather than imagine cats, he imagines torturing children with a pair of pliers and a blowtorch. Is either world preferable?

Smuts thinks that we would obviously prefer world A over world B. He takes this to be true even if we do not assume a consequentialist notion of goodness and harm. There is something inherently morally preferable to world A even if the inhabitant of world B couldn't possible harm anyone.

We might be able to interpret the wrongness of world B on a non-consequentialist moral

stance. From a deontological perspective, we might ask whether the survivor is thinking as a rational agent. A virtue theorist would argue that the denizen of world B is not adopting the attitudes that a moral exemplar should. The problem here is that fictional characters are not real—so how could we construe them as rational agents? On these views, then, I contend that no harm is directly achieved through taking pleasure in fictional suffering.

As Feagin points out, our engagements with fictions free us from typical emotional and moral responsibilities. I agree; the danger of taking pleasure in fictional suffering only arises if a reader adopts a moral judgment about the fictional case and then applies it to comparable real-life things. For example, a viewer might modify his positive evaluation of Voldemort's "Magic is Might" campaign, which condones eugenics and invasive surveillance, and applies it to real-world politics. This is essentially Plato's worry in *The Republic*—artworks do not properly engage our reasoning faculties and so may undermine our moral judgments.

My arguments in this section will be important in the following chapters, in which I discuss the sympathy for the devil phenomenon, the puzzle of imaginative resistance, and the possibility of moral learning from fictions. As we will see, I suspect that much of the confusion surrounding these puzzles arise due to inadequate notions of emotional and moral judgments of fiction (and in general) as well as a confusion the potential harm of experiencing moral wrongness in a fiction.

Chapter 7: Sympathy for the Devil

1. Admiring devils

Imagine reading John Milton's epic poem, *Paradise Lost*. The story begins with a group of fallen rebel angels gathered in their dark, depressing new home. They sit, feeling downcast and defeated, until their stalwart, dashing leader steps forth with words of encouragement and fortitude. This is, of course, Satan, consoling the rebel angels in Hell. As you're reading, it is almost impossible to resist admiring this larger-than-life character who shows such determination and resilience even in the face of certain defeat by their infinitely stronger enemy.

Most readers of *Paradise Lost* likely find Satan an intriguing, admirable, and, let's face it, a sympathetic character. In a strange way, we want him to succeed in his ventures. You know that you are not supposed to admire Satan and his followers—in fact, you should probably feel disgusted by them. However, the poem sometimes makes it difficult to do so. Consider Milton's description of Satan: "He, above the rest/ In shape and gesture proudly eminent,/ Stood like a tower. His form had yet not lost/All her original brightness; nor appeared/ Less than Archangel ruined, and the excess/ Of glory obscured, as when the sun, new risen,/ Looks through the horizontal misty air/ Shorn of his beams…" (Milton, 1667/2007, 18). There is something admirable about the Prince of Darkness' sheltering attitude towards his fellow fallen angels: "Cruel his eye, but cast/ Signs of remorse and passion, to behold/ The fellows of his crime…" (*ibid*, 18). Recall Satan's determination to make the best out of a remarkably bleak situation: "Is this the region, this the soil…this the seat/That we must change for Heaven; this mournful gloom/For that celestial light? Be it so!...The mind is its own place, and in itself/Can make a Heaven of Hell, a

Hell of Heaven" (*ibid*, 9). You have to admire Satan's positive attitude. You cannot help it; to your great surprise (and the consternation of many a Sunday school teacher), you feel sympathy for Satan.

Readers have been fascinated by the portrayals of immoral anti-heroes in fiction long before Milton. Yet this phenomenon, albeit commonplace, raises interesting aesthetic and moral questions. After all, these are characters whom we would likely despise if we met them in the course of our daily lives, but for whom we feel startlingly strong pro-attitudes when encountered in works of fiction— that is, strong positive emotions, thoughts, and feelings. Milton's Satan is likeable and sympathetic. To a certain extent, we want him to succeed in his quest for inner peace.

Following Noël Carroll, let's call this the *sympathy for the devil phenomenon* (SDP) (Carroll 2004). This covers any of our pro-attitudes towards immoral or unlikeable fictional entities including, but not limited to, sympathy. Other pro-attitudes include emotions such as admiration, compassion, empathy, pity, pride, and joy. There is no doubt that this phenomenon occurs on a regular basis. Just consider our pro-attitudes towards *Mad Men*'s Don Draper, *Kill Bill*'s Bride, *Lolita's* Humbert Humbert, or, indeed, *Paradise Lost*'s Satan—immoral characters, all. The question is: *why* do we have the positive feelings and responses towards these characters that we do? Certainly these characters' charming, likeable features do not outweigh their immoral traits and actions. Are our moral feelings so easily swayed?

I will examine and respond to these questions in the following sections. In the previous chapter, I presented an argument for our moral judgments of fictions as integrated with our real life beliefs, values, and emotions. The SDP seems to present a problem for this claim—our responses are often quite different than they would be in real life! I will first consider several potential solutions to the SDP, beginning with Gregory Currie's decidedly distinct attitude-based

simulation theory (ST) approach of sympathy for fictional characters. I will follow that with a critique of Matthew Kieran's *distancing approach* (Kieran 2010) and Carroll's *"best of all characters" view* (Carroll 2004 & 2008). Ultimately, I find each of these views to be inadequate explanations of the SDP. I understand the SDP as a genuine psychological phenomenon that arises in the course of our engagements with fictions (and perhaps in our real lives as well). If that's true, then we should be able to explain the phenomenon in terms of our actual mental tendencies. In this spirit, I will propose an emotional pre-condition for our sympathy for immoral characters: they must elicit the emotion of fascination. In the final section, I will propose my own explanation of the SDP, *the fascinated attention approach.* This approach makes the SDP a matter of narrative content and our emotional response to that content. Thus, it is fully compatible with the SAV.

My view makes up for the inadequacies of the other potential theories. This includes a point that is often not considered when philosophers discuss the SDP: we often feel pro-attitudes towards immoral characters because of their viciousness and other "negative" qualities, and not solely their admirable ones. As Colin Radford pointed out, we sometimes love the rogue *because of* his roguishness (see Smith 1999).

2. Simulated sympathy

Gregory Currie's view draws upon simulation theory (ST) in order to explain the SDP (Currie 1997). We've already encountered this view in chapter 4. Despite Currie's withdrawal from a strong simulative stance in recent work, this view is worthy of consideration because it remains one of the most thorough and articulate examinations of a DAV-based approach to the SDP.

Simulation theory remains a popular theoretical framework used for understanding how we attribute mental states to others and predict their behavior (Currie & Ravenscroft 2002). Like many philosophers and psychologists, Currie understands simulation as "putting oneself in the shoes of another" (Currie 1997, 66). We imagine what it would be like to be in the situation of another person and possess beliefs, desires, and emotions that are similar to theirs. Currie explains:

> I take on, temporarily, the beliefs and desires I assume someone in that situation would start off by having; they become, temporarily my own beliefs and desires. Being, thus temporarily, my own, they work their own effects on my mental economy, having the sorts of impacts on how I feel and what I decide to do that my ordinary, real beliefs and desires have (Currie 1995, 252-253).

When I imagine myself in the position of another person, I imaginarily take on their beliefs, desires, and perspective as closely as possible. We come to understand the experiences and mental lives of others by recreating, as best as possible, their own attitudes in ourselves. From there, we extrapolate information about the target by asking how *we* would feel in a similar situation.

According to ST, the same basic principles apply to our mental engagements with fictional characters (although there is an additional step to simulating fictional characters that I will return to momentarily). When we watch a film, for example, we simulate the mental attitudes of a character. This is typically, but not always, the work's protagonist. We can understand how the character feels and predict what she will do next by imagining what it would be like to be in her position. We can thereby understand what the character experiences: her beliefs, thoughts, emotions, desires, etc. This understanding, in turn, allows us to predict the character's actions, just like in real-life contexts. For example, while reading *Les Miserables* I may imagine what it would be like to be Jean Valjean after he escapes from his 19 year imprisonment working in the French galleys and then runs off with two silver candlesticks from a church. I imagine that I am in Valjean's situation: the relief of freedom, but also destitution of poverty and loneliness. I note that

I feel bitter, angry, guilty, and relieved at the same time. I then extrapolate this information about how I would feel and attribute these feelings to Valjean in his current state.

ST may be able to explain how we attribute mental states to fictional characters and predict their behavior. But can it explain sympathy for fictional 'devils'?

One could potentially take issue with Currie's reliance on ST to explain our care for fictional entities, as we saw in chapter 4. Simulation may not always be required to feel pro-attitudes towards fictional entities. In fact, Carroll has argued that simulation is generally unnecessary for us to understand the mental states of fictional entities (Carroll 2001a & 2008). This is because the narrative does much of the work for us: an author pre-focuses the story so that we emotionally respond in a particular way to the text and a particular character. This means that an author deliberately reveals certain details of a character so that we are able to interpret and understand the story. We are shown positive aspects of a character that causes us to respond positively. Of course, some fictions are emotionally or cognitively *opaque,* so it is quite difficult to understand how a character thinks or feels. It will take more cognitive work to understand them, including understanding the characters' context, past actions, and personality. Generally speaking, however, a narrative can guide our emotional and moral responses to particular characters. Simulating the mental states of a fictional entity is often unnecessary; we can already tell, based on contextual and narrative information, how a character feels without delving into her psyche.

It's possible that simulation is necessary for understanding a fictional entity's perspective and actions. However, it is a further question whether we must simulate the character in order to *care* for them. While Currie makes both claims, it's worth noting that the first claim doesn't entail the second. We may be able to have positive emotional or moral responses for a character without simulating her mental state. For example, we observe Frodo Baggins valiantly offering to take the

Ring to Mordor while watching the *Lord of the Rings* films. We may admire Frodo's bravery even without understanding the mental states involved in his decision. The same works for immoral characters; I watch the self-described serial killer Dexter Morgan save his sister from another murderer. I can admire him for his actions even without delving into his intentions and beliefs for doing so. Emotions about other people in general don't require simulation; for instance, I do not need to understand a friend's current mental states to feel angry or compassionate about her. So why think that they do in the fictional case? This remains unclear on Currie's account. Note that my multi-level appraisal theory of emotions does not require simulation at all.

For the sake of argument, let's grant Currie's claim that we must simulate the mental states of fictional entities in order to care for them so that we can draw out the implications of his view. Assuming that Currie is right, ST can explain the care that we feel for some fictional characters: namely, the ones with positive and admirable qualities, like Frodo. But what about characters who possess many *un*admirable or immoral characteristics? Currie calls this "the problem of personality": I would not normally care about someone like Satan, so why do I seem to during my engagement with Milton's work.

I mentioned above that simulating a fictional entity requires one additional step. Currie argues that we do not *directly* simulate the mental states of the fictional character. Instead, we simulate the mental states of a "hypothetical someone who is reading a factual account of the adventures of [that] fictional character" (Currie 1997, 72). Simulation is transitive; I simulate someone who is simulating a character and who, further, believes that the fiction is real. I am able to understand the mental states of the fictional character by understanding those of the hypothetical someone. This allows me to care for a character like Frodo Baggins; I take on belief-like and emotion-like states as the result of the transitive simulation that gives me access to his mental

states. I care for Frodo because I have access to his many admirable and sympathetic characteristics.

Let's try to make sense of this mysterious "hypothetical someone." There is a vast literature in literary criticism and philosophy of art concerning the presence of an implied narrator: a fictional entity that is not equated with an explicit narrator, actual author, or implied author, who guides the audience through a fictional narrative (see Livingston 1997 and Wilson 1997). I tend be unsympathetic to such views, but I will not take up this argument here. However, the implied narrator is also distinct from the hypothetical someone we simulate. The implicit narrator may not necessarily do any simulating or emotional interpretation of a character. It just presents the fictional story. The hypothetical someone must do some emotional and moral interpretation of a character so that we simulate that character correctly.

There are several other worries with this hypothesis. Following Carroll, I question Currie's claim that we cannot directly simulate the mental states of a fictional character. Note that we can make the same point about the hypothetical someone. Why is it that we can simulate this hypothetical someone, but not the fictional character? This seems especially problematic, since the fictional character, at least, has human-like qualities that we can perceive, but the hypothetical someone doesn't. In fact, there is no direct evidence of a hypothetical someone in the narrative.

Another complication arises here. Simulating a hypothetical someone might help to explain our pro-attitudes towards morally praiseworthy characters like Frodo. However, in order to explain the problem of personality, it is not enough that we simulate the states of the hypothetical someone. The audience must imaginatively take on new versions of beliefs that she currently holds for the duration of her engagement with the fiction. This includes moral beliefs. The reader adopts *pretend* versions of her relatively stable beliefs and preferences (Currie 1997). This allows her to

consider moral propositions that she normally would not. As Currie states: "I imagine myself not merely to be reading fact, but to be someone with an outlook different from my real one" (*ibid,* 73). The greater the difference between our actual moral beliefs and those of the fictional character, the further I would have to adapt my current beliefs and take on new imaginary ones. In order to feel sympathy for Milton's Satan, I would have to adopt the mental beliefs of a radically different kind of person.

One worry here is that it would be quite complicated and inefficient to modify our mental attitudes in the way that Currie suggests. In order for Currie's proposal to work, the viewer would have to quickly and unconsciously switch between different moral viewpoints on multiple occasions as the story progresses. For example, when I watch the film *The Silence of the Lambs,* I simulate a hypothetical someone who simulates the FBI agent-in-training, Clarice Starling. Clarice's moral outlook is generally very close to my own, so I do not need to drastically rearrange my moral beliefs while simulating her. But when I consider Hannibal Lecter, I must adopt an entirely new, psychopathic, hyper-intelligent, and pro-cannibalistic set of beliefs. Consider any of the scenes during which Dr. Lecter and Clarice verbally spar: I may, in the course of one scene, both pity and feel sympathy for Hannibal (consider Dr. Chilton's bone-chilling power plays, how the guards dress him as a monster, how he has not seen the sun in eight years) as well as Clarice (she's actually trying to prevent a death). In order to understand both Dr. Lecter and Clarice's mental states, we would have to switch our simulations back and forth between them rapidly as the scene progresses. A viewer would have to "take on and off" different sets of moral beliefs, one for pro-attitudes for Lecter and another for pro-attitudes for Clarice, since these feelings may often be inconsistent. There is quite a bit of cognitive work going on here, none of which is consciously accessible to the viewer. Indeed, we often find engaging with fictions relaxing, which might be

surprising given the complexity of Currie's model.

There is a further worry here concerning the empirical and phenomenological plausibility of this account. In particular, it is not clear that there is any independent support from social cognition literature that we are required to simulate a hypothetical someone in order to understand nonexistent entities. Currie introduces the hypothetical someone solely to explain the problem of personality. This in itself is not a problem, if the proposal captures something important about our experience of fiction or solves a problem that cannot be solved in any other way. Unfortunately, the simulation view appears to be counterintuitive, or at least phenomenologically inaccurate. It is not clear that any ordinary consumer of fiction would accept that they engage with a story in this way.

Currie could counter that the simulation process occurs unconsciously. That is possible in some cases, but one would think that the simulation must occur consciously at least sometimes. But it is not clear that it does. Furthermore, simulating a hypothetical someone raises a host of further theoretical challenges, some of which I have already mentioned. How could we pick the hypothetical someone that we simulate? What sorts of features would it have? Will everyone simulate the same hypothetical someone, or will there be unique 'someones' for every audience member? Such questions cannot be answered without complicating an already complicated theory.

Complexity is not *in itself* a criticism against a theory, but surely a simpler theory would be preferable if one is available. The main problem with Currie's view is that it makes far too many assumptions concerning the nature of our mental states towards fictions and the ontology of fictional characters. Because of this, Currie (and perhaps some other simulation theorists) is forced to adopt a radical view of our engagement with fictions that does not appeal to our actual experiences. If asked, I suspect that very few readers would accept that they simulate the states of

a hypothetical fictional someone, or even grant that such a thing exists—let alone that they adopt, even for a short time, a radically different moral outlook on the world. Instead, it is more likely that the reader maintains her current beliefs and takes the fiction as an opportunity to explore alternative viewpoints and scenarios that she would not normally encounter.

3. Distancing, pre-focusing & "the best of all characters"

We have questioned Currie's simulation-based approach to the SDP. Now we can explore two other theories that attempt to explain our pro-attitudes towards immoral characters: Kieran's *distancing account* and Carroll's "*best of all characters" approach*.

Kieran's proposal points out that we are able to sympathize with an immoral character because we suppose that the character inhabits a fictional world that is quite different from our own. This fictional world encompasses a different land with different rules, including moral rules. Call this the *distancing approach*. Matthew Kieran argues that imaginative distancing amounts to a psychological distance between an audience and a devilish character (Kieran 2010). We feel free to allow ourselves to feel pity, compassion, and sympathy for someone like Hannibal Lecter or Milton's Satan because of this psychological distance.

"Other-world" distancing may go a long way towards explaining how we are able to become absorbed in a story, for it forces us to realize that the fictional character belongs to a world that is very different than our own. Still, I am unconvinced that this is the right response to the SDP. Satan occupies a very different world from the one I believe myself to live in, and yet I still think that many of the features of that world, including its moral features, track from my world to

that one. So I *should* still find Satan morally repugnant in Milton's story. The other-world distancing also does not seem to work for a realistic film like *The Silence of the Lambs*. This fictional world appears to be just like our world in most ways, including local West Virginian geography, the president of the United States (George H.W. Bush), and the struggles that American women face when they work in a male-dominated field. I do not imagine that Dr. Lecter occupies a distant world from my own; I imagine that he occupies one that is very similar to mine in important respects.

There is another way in which the distancing thesis is found wanting. When we feel sympathy for Milton's Satan, we recognize that this character is physically nonexistent in our world, yet we nevertheless feel profound interest, pity, and compassion for him. The point is that we are generally not *emotionally* distanced from immoral protagonists even if we are in other ways (physically or ontologically). There is no question that we emotionally respond to fictional objects. The problem for Kieran is that we are either *emotionally* distant from fictional worlds and characters (as well as physically and ontologically distant) or, alternatively, we are emotionally close to them despite physical or ontological distance. The SDP shouldn't arise on the former position, but it does on the latter. Yet, we have granted that it surely *does* occur quite often. This should lead us to favor the second position: emotional closeness without physical or ontological closeness. Kieran's position cannot account for this; he does not explain why emotions differ from other kinds of imaginary engagement, why we are imaginatively distant from fictional characters while still emotionally close to them.

Consider another way of understanding distancing, one that is slightly different from Kieran's approach. A fictional world doesn't have to be very *different* from ours in order for us to be *distant* from it. It could be that the reason we feel sympathy for immoral fictional characters is

because we simply do not have to encounter them in our daily lives. I do not have to go to lunch with Hannibal Lecter, or the movies with Satan. Because these characters are not a part of my life, or even my world, I am able to feel some sympathy for them. This implies that we should also feel *some* sympathy for immoral people in our real lives. I accept this possibility and will return to it in the final section. The current problem with this interpretation of the distancing approach is that it doesn't guarantee a positive response to *any* immoral people. Lenin and Ted Bundy aren't a part of my life. I will never encounter them. But this does not guarantee that I will feel sympathy for them. We need an explanation of what triggers our sympathy to begin with.

We have already encountered one way to explain our emotional closeness to an immoral character: emotional prefocusing (Carroll 2008 & Smuts *in prep*). An author may intend for her audience to feel pro-attitudes towards a particular character. This character is often morally corrupt or deviant in some way but, for whatever reason, the author desires for the reader to sympathize with her.

I agree with Carroll that prefocusing is an important part of an explanation of how and why we respond to immoral characters in the ways we do. Our emotional responses to characters are often a matter of how a narrative is constructed and how the details of a story influence our feelings about particular characters. Currie's simulation-based approach did not take advantage of this feature of fictions. However, prefocusing alone does not give us a complete explanation of the SDP. It does not tell us what it is about the immoral character that we admire, or how the narrative must be constructed in order for us to feel sympathy for her. The challenge is to flesh out the notion of pre-focusing to explain these points.

Carroll's own view attempts to do this. He suggests that the reason we feel sympathy for Tony Soprano of *The Sopranos*, Tyrion Lannister of *A Song of Ice and Fire*, or Ethan Edwards in

The Searchers is because, despite their flaws, they are morally better off than the other characters in the fiction. Tony is surrounded by an astonishing array of violent, manipulative, power hungry mobsters. Tyrion is a clever, witty, well-meaning louse with rotten family members. Ethan Edwards is gruff and brutal, but also loyal and in possession of a certain code of honor. So when we search the fictional world for an emotional allegiance, these are the characters we choose. I will refer to this as the *best of all characters solution*.

The morally best individual in a group of bad people often does seem to us like the person with whom we should throw in our allegiance, both in real-life and fictional worlds. Indeed, the best of all characters explanation may explain a great deal of our pro-attitudes for immoral characters.

However, I think that there are two significant problems with this view. First, the "best of all characters" solution doesn't explain why we *ever* have positive responses to an immoral character. We can grant that Tony Soprano is the least bad of a sordid group of mobsters, and so he wins out in gaining our affection. But this does not explain why we feel sympathy for an immoral character to begin with. It's possible that we are deliberately shown features of Tony Soprano that are relatable and positive: his fears of growing old and impotent, his family woes, etc. It's another question whether these glimpses into Tony's psyche override the overwhelmingly abundant immoral actions that he performs—for instance, when Tony brutally attacks his driver on a whim (to use Carroll's example). Surely we should be disgusted by this portrayal of gratuitous violence and turn off the television. And yet, we do not; we continue to watch Tony's exploits season after season, in spite of his immorality. The best of all possible characters explanation cannot account for this.

The second worry is that the best of all characters solution does not apply to all cases.

Perhaps we align ourselves with Tony Soprano because the other mobsters are far more immoral than he is. But there are other fictions that portray morally praiseworthy characters alongside immoral characters, as in *The Silence of the Lambs*. One could argue that we feel sympathy for Dr. Lecter because he allies himself with the morally virtuous Clarice Starling. But does this alliance really explain that sympathy? At best, I think that their relationship explains why we do not *condemn* Lecter the way we would if he was completely unhelpful to Clarice's investigation. It does not explain the pity and compassion that we may feel for him. Or consider a reader's sympathy for Nabokov's Humbert Humbert. Some readers may want HH to succeed in his seduction of the young Lolita. But why? Surely the reader thinks that pedophilia is morally repulsive even if it is presented to us in the guise of a clever, charming protagonist. And HH is surrounded by characters that are far more morally praiseworthy than he is. Finally, think about a story like *Dexter*, in which we feel sympathy for the immoral protagonist despite the fact that he is certainly not the morally best character in his fictional world. We cheer on Dexter in his murderous pursuits even as he faces the brash, morally upstanding Sergeant Doakes or even Dexter's principled sister, Deb. The best of all characters solution cannot explain our sympathy for Dexter.

The distancing and best of all characters explanations both fall short of fully explaining the SDP. Each of the potential solutions suggests that we can explain our sympathy for immoral characters in terms of some kind of emotional manipulation. However, they do not propose a particular emotion involved in shaping our positive responses to immoral characters. Perhaps doing so would help flesh out an explanation of our pro-attitudes towards immoral characters.

4. Fascination for the devil

My view explores sympathy for the devil as a psychological phenomenon. I contend that there is a specific emotional response that underlies our sympathy for immoral characters. There are two points to consider here. First, we should examine the immoral characters' qualities to discover what it is that makes us have such strong positive responses to them. Second, we should consider the actual types of positive responses we have. This will give us a better idea of what kind of explanation the SDP requires.

Recall Milton's description of Satan in *Paradise Lost*. He is physically awesome, intelligent, and deeply loyal to his family and friends. He has attractive mental and physical qualities. The same can be said for other immoral but likeable fictional characters. Humbert Humbert is deliberately portrayed as handsome, clever, and funny. The psychopathic serial killer Dexter Morgan is also handsome, highly intelligent and innovative, and embodies an admirable degree of loyalty to his family. As Carroll and others have noted, positive qualities like these examples are imperative to our sympathy for immoral characters.

I want to push this line even further. Besides these attractive physical and mental traits, Satan, HH, and Dexter are all *exotic*; we do not often encounter people like these in our daily lives, if ever. They are those rambunctious "bad-boys" (to use a hopelessly gendered term; it's an interesting question in itself why most prominent fictional "devils" are men) that charm and allure us with their daring and wild antics. In short, all of these characters possess desirable features alongside their immoral ones. Now imagine a bland, intellectually dull, physically unattractive, run of the mill Satan, HH, or Dexter. Would we still feel pro-attitudes toward such a character? Surely we would not admire them to the same degree, if at all.

Next, consider our actual responses to these characters. We find Satan, HH, and Dexter *intriguing*. We are *curious* to learn more about them. We *pity* and feel *compassion* for them when we learn about the awful things that happened in their past, or when other characters manipulate or exploit them. We are often *glad* when they succeed, and generally *disappointed* when they fail. Perhaps we recognize that our gladness and disappointment are, in some ways, perverse; we take a secret delight in their badness. We know that it might be inappropriate to feel sympathy for an immoral character, but we can't help doing so. Such feelings may be reasonable in the context of the story. There are probably instances in which we have negative emotional responses about these characters. We feel anger or indignation towards an immoral action that appears completely self-interested or entirely callous (e.g. when Hannibal Lecter eats the face off of one of his prison guards) even if we accept some other undoubtedly immoral action that is somehow understandable (plotting to "have Dr. Chilton for dinner").

Synthesizing the previous two points—the nature of the immoral characters and our emotional responses towards them—I want to suggest that underlying our pro-attitudes towards immoral characters is a particular emotional response: *fascination*. I claim that fascination is a specific type of emotion, similar to curiosity and interest, but distinct from them in terms of the objects that elicit it. There are two questions worth asking here: 1) is fascintion actually an emotion? and 2) How can this help us explain the SDP?

If fascination is an emotion-proper, then it has the same features as other emotions. Most philosophers and scientists working on emotions argue that emotions are a kind of evaluation of an object that bears on our well-being (see chapter 5). We feel fear about something that can harm us or those we care about. This includes a rattlesnake basking in the sun next to a path where you and a family member walk; it also includes fear of an upcoming test that. Both objects, the snake

and the test, reflect our interests and well-being.

Theorists also generally argue that emotions have intentional objects: they are *about* something. We do not just feel mad in general; we are angry because of a harm done to us. Many theorists also argue that objects may possess a particular kind of emotionally relevant property. For example, we will generally experience anger when we perceive an object in our environment that has offended us in some way. We experience pride when we take credit for a something that either we, or someone or something with which we identify, has achieved. We feel joy when we realize a goal. These properties can be quite simple (potential harm to you and yours) or complex (taking credit for something that you have achieved) (see Goldie 2000, Lazarus 1991, & Prinz 2004a).

I argue that fascination meets both of these emotional criteria: it is a reflection of an object that serves our interest or well-being, an object that has a particular kind of property. We are fascinated about particular objects that interest us in the right way; we are not just fascinated vaguely or in general. There is a certain phenomenological, qualitative feeling that we associate with fascination. There is a kind of absorption or interest that we have towards fascinated things. This is important, since we generally think that emotions elicit feelings. Fascination meets this criteria as well.

We also need to know what types of objects generally elicit fascination. I argue that fascinating objects typically have three different features. First, they must be a *curiosity*; the object must be unusual, unique, different, or exotic in some way. It must be something out of our ordinary experience. Second, the object must be *attractive*. This could mean physically attractiven in the case of people (James Dean as Jim Stark in *Rebel without a Cause* or Angelina Jolie as Lisa Rowe in *Girl, Interrupted)* or concrete objects (a sculpture or building), mentally attractive in the case of

people characterized by their cognitive abilities (I would hesitate to deem Anthony Hopkins' Hannibal Lecter a standard of male beauty. Rather, it is Lecter's intellectual brilliance that we find attractive). Or perhaps the object is oddly agreeable in some way, as in the case of a fascinating topic that we find worth pursuing. Lastly, fascinating objects are cognitively *interesting*. They inspire us to learn more about them, and we think that by doing so there will be some kind of cognitive payoff; we will gain some new experience, information, or perspective on the world.

There is no set limit to the types of things we can find fascinating. We can be fascinated by a philosophical idea that we find original, important, useful, or astute. We can be fascinated by artworks: a musical piece, painting, or story. Consider Gerhard Richter's painting, *Betty*. You might be struck by the beauty of the painting and be curious to know more about the artist and the subject: who is that girl? Why is her face hidden from the viewer? Why did Richter paint her in that soft, muted way? Although *Betty* is a representational painting, one could be equally fascinated by an abstract work by Rothko, Pollack, or Reinhardt. Lastly, and most importantly for our purposes, we can be fascinated by *people* that we find unusual, exotic, and attractive. Who is that stranger at the café? Where is she going? What is that accent?

In short, I argue that fascinating things are attractive, interesting curiosities. Generally speaking, these are the kinds of objects that we will be fascinated by; if an object lacks these characteristics, then we probably will not be fascinated by it. Note that each feature can be broadly construed so as to apply to a wide variety of cases. This is not a defect of the characterization, but rather a virtue. Consider the imprecision of our characterization of anger-inducing objects. We are angered by things that we find offensive. This characterization does not tell us which people or actions specifically cause us to experience anger. We can tell a similar story about fear: we fear objects that could potentially harm us, but this leaves open what those objects actually *are*.

Like anger and fear, I leave my characterization of fascination open-ended, yet precise enough to understand what we typically characterize as fascination-inducing. Not everyone will find the same people or objects fascinating, just as not everyone will find the same things attractive, anger-inducing, or scary.

I argue that fascination, so understood, is the key to understanding our pro-attitudes towards immoral, fictional characters (see also Smith 1999). Each of the immoral characters I have discussed so far fit my description of an attractive, interesting curiosity. Take the Count of Monte Cristo: a revenge-driven, almost cruel, mysterious and charming noble. He is surely not the kind of person we encounter on a regular basis! Perhaps we think that by taking an interest in this character we can expand our folk psychology to include the vengeful and obsessive mindset the Count represents. Fascination is the pre-condition for our sympathy towards immoral characters. Generally speaking, we need to be fascinated by immoral characters before we feel sympathy for them. Fascination is achieved by how the character is portrayed in the narrative as possessing the exotic and curious traits that I have described. Once this is achieved, other aspects of the narrative will cause us to feel sympathy for them.

This is not to say that fascination *is* sympathy. We can feel sympathy (in the narrow sense or wide sense I have been employing) for someone with necessarily being fascinated by them. For instance, I can feel sympathy for a close relative who, after many years of close contact, I do not feel fascinated by. I can also be fascinated by things for which I do not feel sympathy, as we will see in the next section.

The fascination approach seems to explain our pro-attitudes better than the other potential solutions that I have considered. But is it enough to explain the SDP? The fact that an immoral character is fascinating may not be enough to generate pro-attitudes towards them. There is still a

missing ingredient in our account of the SDP.

5. The fascinated attention approach

Carroll denies that fascination with a character is enough to explain our pro-attitudes towards immoral characters. It may be that we are often fascinated by Tony Soprano, but this fascination does not explain our *sympathy* for them. There are many characters that we might find interesting in *The Sopranos,* for instance, but we only feel sympathy for Tony. It is only through prefocusing and the best of all characters vantage point, Carroll argues, that we feel pro-attitudes towards immoral characters.

One interesting implication of my view is that we will often be fascinated by fictional immoral characters, as well as real-life immoral people. Consider Truman Capote's novel, *In Cold Blood,* and Werner Herzog's documentary, *Into the Abyss*. Both of these narratives portray serial killers in a highly sympathetic light. We feel sympathy for these serial killers partly because we are fascinated by the narratives' portrayal of violence and "evil" personalities.

But fascination alone will not always engender sympathy. Consider contemporary readers' continued fascination with historical figures like Adolf Hitler, Mao Zedong, and Joseph Stalin, three people who are likely responsible for the greatest abuses of human rights in all history. Surely they don't deserve our positive responses. And yet these leaders' biographies continue to be popular. I would doubt, though, that the people who read these biographies would approve of or feel strong sympathy for them, like we do for Satan in Milton's tale. So what is the difference

between the *In Cold Blood*-type cases and these biographies? It cannot be that the former are simply told in a narrative and the latter are not. Biographies are, after all, a *kind* of narrative.

I argued in the previous section that a narrative draws out attractive, exotic features of characters to make them fascinating to us. This is true for morally blameworthy and praiseworthy characters. But our emotional and moral allegiances are not won solely based on what we are *shown* about someone. Our pro-attitudes are also shaped by what we are *not* shown about a character. What would it take for us to feel sympathy for a real-life serial killer, such as Ted Bundy? I think that it would require a particular kind of narrative that utilizes a particular kind of pre-focusing. The narrative must attract an audience's attention towards the morally praiseworthy features of a person or character and a shift of their attention away from their morally blameworthy features.

Let's consider an example. We saw in chapter 6 that the *Harry Potter* series presents us with a paradigm immoral antagonist: Lord Voldemort. Voldemort is *supposed* to be a highly unsympathetic character. Almost everything we know about him confirms this point. Voldemort kills innocent people, including children, he abuses his followers, he is disturbingly racist, and he fits every criterion for antisocial personality disorder. There are few instances in the story in which one might feel sorry for Voldemort. Still, we may be fascinated by Lord Voldemort. We are curious about his past and his motivations; we are in awe of his magical power. But we do not feel sympathy for him. We do not wish for Voldemort's "Magic is Might" plot to succeed, and we do not feel compassion for him when he meets his fate at the end of the series.

We do not feel full-blown sympathy for Voldemort because we are also shown features of him that override his potentially pitiable qualities (for instance, his difficult childhood and loneliness). Imagine a retelling of the *Harry Potter* series from Voldemort's perspective. There are

several examples of this: *Wicked* (wherein audiences are retold *The Wizard of Oz* from the Wicked Witch of the West's perspective), *Grendel* (a retelling of *Beowulf* from the monster's point of view), and *Wide Sargasso Sea* (the story of *Jane Eyre*'s "mad lady in the attic"). In our retelling of *Harry Potter*, we would learn more about Voldemort's pitiable childhood and his justification for why he began his life of crime and hatred. The story would brush over some of the more sordid details of his violent behavior and highlight his fascinating characteristics. We would gain new perspectives of the more likeable characters in the original *Harry Potter* series, such as Dumbledore, McGonagall, and even Harry Potter himself.

The retelling may be successful in eliciting sympathy for the Dark Lord if we are shown the pitiable aspects of Voldemort and are kept from considering others. We can generalize this example. An audience must be given certain information in a story that will create pro-attitudes towards the immoral character as well as distract their attention from the less admirable aspects of his/her behavior and personality.

We now have the means to fully explain the SDP. Not only must we be fascinated by a devilish character, we generally must also be shown particular features of a character that deliberately focuses our attention away from her immoral behaviors or traits, and highlight the fascinating characteristics. I call this the *fascinated attention approach*. Its two components (fascination and attending to certain features of the character) are both in place when we feel sympathy for an immoral character. This is a psychological claim rather than a metaphysical one; fascination and prefocused attention may not always be necessary or sufficient for the SDP to occur. I argue, however, that they generally *are*. If we think of the SDP in terms of our emotional responses towards a character, then we can understand immoral characters with fascinating features as the kinds of emotional objects towards which we would respond in our real lives. Not

everyone will emotionally respond to the same objects, both in the case of fascination as well as anger, sorrow, or joy.

The *fascinated attention approach* makes up for the shortcomings of the other explanations of the SDP that we have discussed. My view highlights a defect in the distancing explanation: that view is contradicted by the very phenomena of SDP, which is in part generated as a result of our emotional *identification* with the immoral character, rather than our distance from it. Furthermore, fascination approach tells us what it is about a character that makes us feel pro-attitudes towards her, and so improves upon the pre-focusing model. Finally, my view also makes up for the defects of the best of all characters solution. It is not just that Tony Soprano has some positive qualities that outweigh his negative ones, or that he is better than other bad-to-the-bone characters. Indeed, we can be fascinated by Tony *because of*—not in spite of—his negative features. His portrayal of an immoral mobster is unusual, intriguing, and attractive: we do not normally experience someone like Tony Soprano, and so we may have to gain a new perspective of a new character type by studying him.

Additionally, the fascinated attention approach can also explain allegiances that might be considered "incorrect" in the fiction—for example, compassion for Nurse Ratched in *One Flew Over the Cuckoo's Nest* or the Joker in *The Dark Knight*. Some people may be fascinated by these characters while others are not; the former may be inclined to feel pro-attitudes towards characters in a way that isn't intended by the author. The fascinated attention approach also explains why we often feel sympathy for multiple characters in a fiction, including ones that have conflicting aims. This was the case with Dexter Morgan and his sister Deborah Morgan. The best of all characters solution has difficulty accounting for these multiple allegiances.

Clearly prefocusing helps in terms of creating a fascinating immoral character. Aesthetic

devices like music, point of view, editing, narration, etc. all contribute to the positive response we have for antiheroes and villains. It is the task of the artist to create a work with a fascinating character who holds our interest. We would not care that much for Tony if it weren't for his fascinating qualities. One might go so far as to say that portraying a fascinating character is the *raison d'etre* for a great many fictional narratives.

This last point raises a question as to the moral significance of our pro-attitudes towards immoral characters. It might make sense that we have pro-attitudes towards immoral characters. But are such emotions *appropriate*? Should we be held morally blameworthy for having pro-attitudes towards immoral characters? Perhaps an answer to this depends on what one means by blameworthy. If we mean that our fascination is potentially *harmful*, then I would reject this claim. Generally speaking we do not harm actual people by feeling sympathy for immoral characters, at least, not directly (see chapter 6).

Furthermore, *contra* Currie, my guess is that we often have pro-attitudes towards immoral characters while at *the same time* condemning many of their immoral actions. We do not radically shift our moral values for the duration of the fiction or imagine ourselves in an immoral character's position. I pity Hannibal Lecter even while I am repulsed by his brutal murder of the unsuspecting ambulance drivers. I do not accept or condone Lecter's immoral actions. Rather, I am fascinated by this character even as I condemn his actions. It's even likely that I'd be *less* fascinated if Lecter was merely a morally praiseworthy, brilliant criminologist instead of a sociopathic cannibal.

Importantly, unlike Currie's approach, my view is fully compatible with the SAV; nowhere along the way have I appealed to unique or distinctive mental states involved in our sympathy for immoral characters. In fact, I have suggested that the SDP is not uniquely fictional.

Near the end of his discussion of our pro-attitudes towards anti-heroes, Carroll states that

sympathy comes cheap in real-life, but must be earned in a fiction. I disagree. I suspect that we will sympathize with just about anyone on a screen, a stage, or in book that is presented to us for any significant length of time. We do not even need much information about the character. Consider our sympathy for Lola in *Run, Lola, Run*: who is this person? Why is she in this strange situation? Why does she keep making such bad decisions? We care for her even before we learn about how much she loves her boyfriend (her most redeeming quality). It does not matter that we know very little about this character or that we disagree with her choices. We like her anyway. I would go so far as to say that the character does not need to be portrayed in a positive light, like Lola is. Fascination is cheap. If we are shown an interesting immoral character going about his business, chances are we will have some kind of positive attitude towards him (like HH). That's just how we experience fictions.

Chapter 8: The Puzzle of Imaginative Resistance

1. Resisting Humbert Humbert

The sympathy for the devil phenomenon captures one facet of our moral experiences of fictions. We often feel pro-attitudes towards fictional characters that we would dislike or be disgusted by in our real lives. One of my central claims in the last chapter was that we may feel sympathy for these immoral people in real life if their stories were framed in the right way. We could be fascinated by them and attend to certain aspect of their personality and history, while ignoring others.

There are also cases in which we take a profound dislike to immoral fictional characters even when an author may intend for us to admire them. Take the protagonist and narrator of Nabokov's *Lolita*, Humbert Humbert. As HH tells his story, we learn of his compulsory attraction to young "nymphets," and especially to the spritely Lolita. He propounds detailed arguments in favor of pedophilia: it has been practiced through the ages, in more "civilized" times than ours, by royalty, aristocracy, and many of the great artists; he is more interested in the spiritual aspects of nymphets than the physical; he cannot help how he feels. Some readers probably remain unconvinced of HH's moral innocence despite his arguments. Others might feel torn by HH: we admonish his tyrannical treatment of the young girl, yet at the same time feel betrayed right along with him when Lolita runs away. HH is remarkably funny and clever and his storytelling is evidence of genius.

Yet, for every quip, literary reference, and Quilty clue, there is some horrifying detail about

Lolita's desolate condition: her quiet sobs and winces, her raging outbursts, HH's matter of fact admission that she "had absolutely nowhere else to go" (Nabokov 1970, 144). Part of what makes *Lolita* work is that some readers often do feel some sympathy for HH. Some readers, though, will probably never feel sympathy for him. Perhaps this is an aesthetic flaw with the text; Nabokov hasn't done enough to motivate HH's perspective and, further, no text should ever portray pedophilia in a positive light. Alternatively, the reader herself may make a mistake by not feeling sympathy for HH, thereby missing out on an important cognitive point of the narrative.[18] For whatever reason, many readers resist Humbert Humbert.

Generalizing this case, some philosophers have suggested that readers will resist characters whose immoral values and actions are sufficiently dissimilar from their own. These characters stretch the limits of our literary imagination, but imagining these immoral characters might be required for the narrative to work. As HH admonishes his readers before a pivotal scene: "Please, reader: no matter your exasperation with the tenderhearted, morbidly sensitive, infinitely circumspect hero of my book, do not skip these essential pages! Imagine me; I shall not exist if you do not imagine me…" (*ibid*, 131).

Near the conclusion of "Of the Standard of Taste," David Hume makes several remarks on the moral status of fiction that are particularly relevant for us:

> [Where] the ideas of morality and decency alter from one age to another, and where vicious manners are described, without being marked with the proper characters of blame and disapprobation; this must be allowed to disfigure the poem, and to be a real deformity. I cannot, nor is it proper that I should, enter into such sentiments; and however I may excuse the poet, on account of the manners of his age, I never can relish the composition. ..And whatever indulgence we may give to the writer on account of his prejudices, we cannot prevail on ourself to enter into his

[18] In a BBC television interview, Nabokov declared *Lolita* his favorite book (P.D. Smith 1962). In a *Playboy* interview, he joked: "I am probably responsible for the odd fact that people don't seem to name their daughters Lolita any more. I have heard of young female poodles being given that name since 1956, but of no human beings"(Toffler 1964).

> sentiments, or bear an affection to characters, which we plainly discover to be blameable.
>
> The case is not the same with moral principles, as with speculative opinions of any kind.
> These are in continual flux and revolution. The son embraces a different system from the father. Nay, there scarcely is any man, who can boast of great constancy and uniformity in this particular. Whatever speculative errors may be found in the polite writings of any age or country, they detract but little from the value of those compositions. There needs but a certain turn of thought or imagination to make us enter into all the opinions, when then prevailed, and relish the sentiments or conclusions derived from them. But a very violent effort is requisite to change our judgment of manners, and excite sentiments of approbation or blame, love or hatred, different from those to which the mind from long custom has been familiarized (Hume 1757/1994, 90-91).

So began the discourse on *the puzzle of imaginative resistance*: although we may be willing to accept factual or metaphysical discrepancies in a fiction, we may be loathe to accept deviant moral values and practices that are not treated positively by the work.

Hume's comments actually highlight several discrete puzzles, as others have pointed out (Walton 2006 & Weatherson 2004). The first is the *aesthetic puzzle*: if an artwork in some way embodies moral defects, do those defects detract from the aesthetic value of the work? Kendall Walton believes that this puzzle may be only indirectly related to moral resistance (Walton 2006).

Second, the *fictionality puzzle* states that: "we easily accept that princes become frogs, or that people travel in time, in the world of a story, even, sometimes, that blatant contradictions are fictions. But we balk…at interpretations of stories of other fictions according to which it is fictional that (absent extraordinary circumstances) female infanticide is right and proper…or that a dumb knock-knock joke is actually hilarious" (*ibid,* 140). The fictionality puzzle concerns any sort of value judgments, not just moral ones. People may often deny that a value that they reject in the real world is correct in the fictional world (or vice versa). So we may refuse to accept that the dumb knock-knock joke could possibly be funny, even in a fictional world. We may also be unable

to accept that female infanticide is morally permissible in another world because we don't believe that it is in ours.

The last part of Hume's worry, the *imaginative puzzle,* does not concern what is or is not fictional, but rather what we can or can't imagine *at all*. We might be able to imagine a situation in which female infanticide is morally acceptable, even if we do not accept that it could ever be fictionally true that it is. Alternatively, we might not even be able to imagine that female infanticide is morally acceptable. We are unable to wrap our heads around the moral acceptability of female infanticide, so to speak. The fictionality puzzle concerns what we are able to *accept as true* in the world of the work. The imaginative puzzle concerns the limits of our imagination.

In this chapter, I will offer ways to reject each of the puzzles of imaginative resistance. I will begin with the aesthetic puzzle. While my SAV does not necessarily favor one particular solution to the aesthetic puzzle, I do think that it narrows the playing field. I will then turn to the fictionality and imaginative puzzles. Traditionally, these puzzles have been kept distinct from the aesthetic one. I think that this is a mistake. My response to the puzzle of imaginative resistance is threefold. First, I deny that the phenomenon of resistance is as robust as some philosophers seem to think it is. I will present several supposed cases of resistance and why they should not be considered resistance in the sense others have argued. Second, I will compare the puzzle of imaginative resistance to the SDP and attempt to show that supposed cases of imaginative resistance are actually cases of *emotional* resistance. Finally, I will argue that some cases in which we seem to resist fictions stems from an aesthetic flaw in the work or a failure on the part of an audience to grasp a work's cognitive value.

The puzzle of imaginative resistance might seem a bit innocuous after our lengthy discussion of the sympathy for the devil phenomenon (SDP); if we can get ourselves to feel

sympathy for devils, cannibals, and psychopathic serial killers, then surely we can get ourselves to consider the possibility of deviant moral codes. I argued in the previous chapter that we typically maintain our own moral beliefs and values throughout the duration of our engagements with fictions. We undergo pro-attitudes towards immoral characters *in spite of* our moral values, because of our fascination with them. The difference between the puzzle of imaginative resistance and the SDP is that, for the latter, the work in question requires that we consider a deviant moral outlook *and* condone it.

2. The aesthetic puzzle

I will begin my discussion of the aesthetic puzzle with a historical anecdote about the Whistler vs. Ruskin libel case of 1877. This case illustrates a turning point in the theorizing about art's purpose. Indeed, the clash of the devout and revered critic, John Ruskin, against the arrogant and witty painter, James McNeill Whistler, was less a legal case than an aesthetic one. Ruskin stood for the established order that held that paintings should represent nature and serve a moral purpose. He stated that "…fine art had, and could have, but three functions: the enforcing of the religious sentiments of men, the perfecting of their ethical state, and the doing them material service" (Ruskin 1870/2006). He found these qualities lacking in Whistler's controversial painting *Nocturne in Black and Gold: The Falling Rocket* (1875). Ruskin denounced the *Nocturne* in the July 1877 issue of his series of letters, *For Clavigera*. He stated:

> 'For Mr. Whistler's own sake, no less than for the protection of the purchaser, Sir Coutts Lindsay ought not to have admitted works into the gallery in which the ill-

> educated conceit of the artist so nearly approached the aspect of willful imposture. I have seen, and heard, much of cockney impudence before now; but never expected to hear a coxcomb ask two hundred guineas for flinging a pot of paint in the public's face' (Whistler 1890/1967, 3).

Whistler then sued for libel. During the trial, the art critic Edward Burne-Jones attacked Whistler's painting on the grounds that it has "fine color" but lacked meaning, and "it would be impossible to call it a serious work of art" (*ibid*, 14-15).

Ruskin and many of the other critics who denounced the *Nocturne* did not accept Whistler's rejection of their moralistic aesthetic standard. Whistler promoted the view that we should value "art for art's sake," distinct from the moral, material, religious, and social obligations that Ruskin believed validate a work. Whistler's aesthete position marks one extreme of the range of responses one can make to the aesthetic puzzle: art comprises an autonomous realm from social, moral, and practical considerations (Carroll 2001a). Call this view *aestheticism*. Whistler's aestheticism was highly influential in years to come; Oscar Wilde and art theorists such as Clive Bell and Clement Greenberg all adopted versions of it (Bell 1914/2003, Greenberg 1940/2003).

Ruskin's position lies on the other end of the spectrum: a moral defect in a work of art is always an aesthetic defect. Call this view *ethicism*. But Ruskin's position goes to an even further extreme: if a work of art lacks some display of positive moral values—as opposed to positively presenting immoral values—then it is aesthetically defective. Plato seems to have held a similar view. In Book X of *The Republic*, Socrates argues that mimetic artwork is inherently morally corrupt, since mimesis is removed from reality—it does not pertain to the truth. But even an abstract (non-mimetic) work like the *Nocturne* would be considered dangerous for encouraging the movements of the baser parts of the human soul: the appetite and spirit (Plato 1985).

More recently, Berys Gaut (2007) has defended a position on the relation of art and

morality that follows Plato and Ruskin in spirit. Gaut argues that there is an inherent connection between aesthetic and moral value. A moral flaw makes the work of lesser aesthetic value; a moral *virtue* in a work makes it of greater aesthetic value. To support this position, Gaut notes that we generally accept that cognitive features of a work are aesthetically praiseworthy. For example, we consider accurate portrayals of emotions and decision making, and interesting intellectual puzzles to be part of the aesthetically good-making features of a work. The problem is that portrayals of immorality are a kind of cognitive defect because it asks the audience to make moral mistakes. Gaut argues that it is *necessarily* an aesthetic failure when a work misrepresents moral actions or principles, solicits perverse moral responses, or condemns morally praiseworthy actions (Kieran 2006). This is because such portrayals involve a misunderstanding of how we should treat moral actions or principles. They get the audience make poor moral judgments.

I will return to Gaut's ethicist position in §5, when I argue that portrayals of immorality do not necessarily constitute a cognitive defect. For now, we should not that there are a variety of nuanced intermediate responses to the aesthetic puzzle, between Whistler's aestheticism and Gaut's ethicism. One position maintains that aesthetic considerations are separate from moral considerations in *some* circumstances, but not all. Carroll describes this view as *moderate autonomism*. An artwork may be critiqued both in terms of its moral and aesthetic characteristics, but its aesthetic features are distinct from its moral ones (Carroll 2001b). It may be appropriate to morally evaluate an artwork like the *Nocturne*. But such evaluations do not bear on the painting's aesthetic value. The same can be said about fictional narratives that display moral failings. A critic may praise *Lolita* for its displays of literary genius while at the same time condemn the work for its positive portrayal of an immoral character.

Carroll's own position, *moderate moralism*, states that certain genres of narrative fictions

can be evaluated in terms of its moral features, and the work's moral features may contribute to its aesthetic evaluation (*ibid*, 299). This is the case when the moral features of an artwork are paramount to our understanding the work as a whole. We saw in chapters 5 and 6 that audiences often need to apply their own emotional and moral values to artworks in order to adequately respond to a fictional narrative. Moderate moralism captures this point; it would be an aesthetic flaw in some artworks for them to portray immoral actions or principles, since doing so would render the audience incapable of adequately interpreting the narrative. Carroll cites the films *Natural Born Killers* and *Schindler's List* as aesthetically defective because of moral flaws—the former because it "advertises itself as a meditation on violence, but it neither affords a consistent emotional stance on serial killing, nor delivers its promised insight on the relation of serial killing to the media…" (Carroll 2001a, 289), the latter because is excessively engages our moral emotions to the point of sentimentality (*ibid*, 290).

The last position I will consider is *immoralism*. In contrast to the ethicists and moderate moralists, immoralists argue that positive portrayals of immorality can sometimes constitute an aesthetic *virtue* (Kieran 2006, Feagin 2010). Gaut argues that ethicism arises out of the recognition that the cognitive virtues of artworks also constitute aesthetic virtues. Immoralists argue the exact opposite. It may be an aesthetic virtue for a work to display a moral defect because doing so has cognitive value.

Susan Feagin (2010) uses *Gone with the Wind* to promote a version of immoralism. She argues that the moral insensitivity of a work may enhance its aesthetic features—in this case, that positive portrayals of slavery may actually comprise a cognitive virtue and so an aesthetic one as well. Both the novel and film adaptation of *Gone with the Wind* are generally thought to be classic examples of imaginative resistance. We resist the positive portrayal of slavery and enslaved

people. However, Feagin argues that doing so makes us miss out on the moral message of the film. This message is revealed when Ashley Wilkes (arguably the moral compass of the story), decides that he will free his slaves and turn away from the old way of life. Feagin states: "If the would-be appreciators are so morally repulsed by the film's introductory scenes that they are unable to engage with the story, they will not be in a position to appreciate how those appearances are woven into the rich character studies… that emerge out of the epic nature of the film's narrative" (*ibid,* 28). Thus, *Gone with the Wind*'s supposed moral failings might actually be much more complex than is typically assumed. The narrative could even serve as an opportunity for moral learning. Moreover, the narrative's cognitive virtues might also constitute aesthetic merits; the novel and film are aesthetically better because its positive portrayals of immorality make some cognitive point.

I will also argue for a version of immoralism in what follows. Against ethicism and moderate moralism, positive portrayals of immorality may sometimes be aesthetic virtues. My view also rules out aestheticism. Art and morality do not comprise completely separate evaluative realms. We cannot help but treat artworks morally and we sometimes *should* evaluate the moral features of a work for their aesthetic benefits or defects. However, the moral features of artworks may not always bear on a work's aesthetic qualities. We may respond negatively to a positive portrayal of an immoral action and so judge the artwork as morally flawed even while recognizing that the work is aesthetically good *in spite of* the moral flaw. So my immoralist position is compatible with moderate autonomism.

I agree with Carroll that sometimes our emotional and moral capacities are sometimes required for audiences to adequately interpret an artwork. However, Carroll's moderate moralism doesn't spell out the cases in which a moral defect constitutes an aesthetic one; he contends that

this will likely vary case by case (Carroll 2001a). Kieran (2006) thus contends that moderate moralism incomplete. Carroll's theory would collapse into ethicism once we stipulate the conditions under which a moral defects constitutes an aesthetic defect. Furthermore, moderate moralism cannot account for the advantages of portraying immoral values, like immoralism can.

Responding to the aesthetic puzzle doesn't seem to help us with the most challenging aspects of the puzzle of imaginative resistance—namely, why it is that we often seem to resist immoral characters and scenarios. Most philosophers treat the aesthetic puzzle separately from imaginative resistance. Whether or not an immoral feature of an artwork is also an aesthetic defect is irrelevant to the question of how we can imagine a world in which an immoral action is acceptable.

In contrast, I think that an immoralist solution to the aesthetic puzzle plays an important role in answering the other two puzzles. Our resistance to immoral characters or scenarios may sometimes be the result of the way a story is told and what information it provides about an immoral character or action. One could argue, then, that some resistance we experience is an aesthetic failing of a work, because the narrative is incomplete or doesn't adequately convey the cognitive significance of the positive portrayal of immorality. Other times, resistance may be due to the audience who fails to adequately appreciated a point the film is making.

I will now discuss the fictionality and imaginative puzzles. As I mentioned in the introduction to this chapter, there are three parts to my response to these puzzles. I will begin by attempting to undermine the intuitive motivation behind the puzzles. Most philosophers working in the resistance literature seem to think that imaginative resistance is a robust phenomenon. In the following section, I will argue that many of the examples of imaginative resistance that are typically appealed to are not, in fact, genuine cases of resistance.

3. Evidence for the puzzle of imaginative resistance

Fictions often present us with unusual or seemingly impossible situations: amazingly coincidental romantic comedies, zombie stories, time-travel, etc. Audiences usually take the bizarre or impossible features of a story in stride. We often do not question historical inaccuracies or anachronisms and we delight in physical impossibilities, like huge explosions in space or Star Trek transporters that can move physical objects faster than the speed of light. There are exceptions, of course. Consider the heated debates over historical television programs, like the History Channel's *Vikings*, which take artistic liberties with historical facts, or complaints about scientific mistakes in fictions, like Alfonso Cuarón's film *Gravity*. These audiences resist the fictional world because of historical or scientific inaccuracies. These mistakes turn out to constitute aesthetic failings as well—for some people, anyway. In general, though, I think that audiences tacitly apply a principle of charity to fictional worlds. Scientific and historical discrepancies can be explained away because the fictional world is distinct from the actual one. An audience might miss a work's cognitive or aesthetic qualities if they are uncharitable in this respect.

We need to know whether there are any limits to our artistic imagination, at least in principle. Here are several examples that seem to push this limit to the extreme.

The first test case comes from obvious contradictions. This might include something as simple as a round-square. Other inconsistencies can be far more complex. In "Sylvan's Box," Graham Priest tells a story in which a character discovers a box that belonged to a deceased

friend—a box which is both empty *and* occupied (Priest 1997). The story makes up approximately two-thirds of the article. It is arguable that the background narrative is required in order for anyone to buy into Priest's contention about the inconsistency of the box, although my guess is that some readers still will not be able to imagine the empty-occupied box (especially if they have previous metaphysical commitments!). Even in cases where we accept obvious contradictions, inconsistencies like these may be impossible to imagine *imagistically*. In other words, we cannot picture them in our mind even if we can understand what it might mean for a box to be both occupied and unoccupied.

Next, consider *conceptual impossibilities*. In "Coulda, Woulda, Shoulda," Stephen Yablo tells a hypothetical children's story: "They flopped down beneath the great maple. One more item to find, and yet the game seemed lost. Hang on, Sally said. It's staring us in the face. This is a *maple* tree we're under. She grabbed a five-fingered leaf. Here was the oval they needed! They ran off to claim their prize" (Yablo 2002, 485). A five-sided oval is conceptually impossible and, again, perhaps impossible to imagistically imagine without undergoing duck/rabbit-like shifts from one shape to another.

Third, there are *metaphysical* impossibilities. Brian Weatherson's cites the "Restaurant at the End of the World" in the second installment of the *The Hitchhiker's Guide to the Galaxy*:

> The Restaurant at the End of the Universe is one of the most extraordinary ventures in the entire history of catering.
>
> It is built on the fragmented remains of an eventually ruined planet which is enclosed in a vast time bubble and projected forward in time to the precise moment of the End of the Universe.
>
> This is, many would say, impossible… (Adams 1980/2002, 213; quoted in Weatherson 2004, 9).

In fact, almost every futuristic sci-fi—*Star Trek*, *Star Wars*, *Firefly*, *Battlestar Galactica*—features metaphysical impossibilities of some kind, such as space-travel or, indeed, a restaurant surviving

the end of the universe. While it may be difficult to understand these impossibilities, we do not generally resist them. Most viewers probably do not question warp-speed space travel in fictions. They simply go along with this impossibility because it is stipulated in the story. Our world and the fictional world of the fiction differ in this important respect. We are willing to take for granted whatever technological or physical discrepancies that allow for this.

Finally, and most importantly for our purposes, we may resist deviant evaluative claims, including moral claims. Yablo argues that we resist deviant aesthetic evaluations like the following:

> All eyes were on the twin Chevy 4x4's as they pushed purposefully through the mud. Expectations were high; last year's blood bath death match of doom had been exhilarating and profound, and this year's promised to be even better. The crowd went quiet as special musical guests ZZ Top began to lay down their sonorous rhythms. The scene was marred only by the awkwardly setting sun (Yablo, 485).

Yablo contends that a typical reader would resist the claim that a "monster truck death match" could be profound, yet the sunset could somehow be aesthetically awkward. Certainly we can think of exceptions to Yablo's resistance. Some people might admire monster trucks. They may think that there is something primal and profound about a monster truck death match and something rather ho-hum about a sunset. Such audiences would not resist Yablo's short fiction, even if others find it strange or implausible.

There are other evaluative claims that we might resist. Consider Walton's dumb knock-knock joke: "Knock, knock. Who's there? Robin. Robin who? Robbin' you! Stick 'em up!" (Walton 1994, 43). Walton argues that we cannot bring ourselves to think that this dumb knock-knock joke is funny. We cannot imagine that it is, even in a fictional world. Again, I do not think that this is quite fair. Just because I do not find this joke funny does not mean that I cannot *imagine* that it is. I probably would have found this joke funny when I was five.

This brings us to moral claims. Can we imagine a morally deviant situation and, further, fictionally accept its truth? Take Walton's female infanticide example: "In killing her baby, Giselda did the right thing; after all, it was a girl" (*ibid*, 38). Walton suggests that a reader will probably not be able to imagine the moral acceptability of female infanticide. The reader might put down her book and walk away or do something else to prevent her from engaging with the immoral statement. She certainly will not accept that female infanticide is morally acceptable in the fictional world.

I do not find any of the previous examples of contradictions, impossibilities, and evaluative anomalies to be strong evidence in support of the resistance phenomenon. As I've pointed out, many people will not resist deviant evaluations or impossibilities and we can rather easily imagine not resisting them. But some philosophers have taken the moral cases to be particularly problematic. Why would this be?

4. Responding to the puzzles

There are three distinctions that we must make when considering the above examples. First, we must be careful to distinguish between the imaginative puzzle and the fictionality puzzle. The former concerns what we do or do not imagine, regardless of whether it is true in a fiction. We resist imagining certain scenarios and propositions, including non-evaluative propositions and scenarios. Our task is to determine whether this failing is the result of something in the story or something about the subject. In contrast, the fictionality puzzle concerns our resistance to

accepting that a deviant evaluative claim *is true in a fictional world*. This can include deviant aesthetic and humorous claims, as well as moral ones. In many cases it will be important for us to keep these two puzzles separate. As we will see, though, it is not always easy do so.

The second important distinction concerns the *type* of imagining involved in our resistance to fictions. As the above examples show, we may find ourselves resisting to *imagistically* imagine a scenario (difficulty in visualizing it), attitudinally imagine it (take the proposition as true in the fiction), or constructively imagine it (create further propositions or scenarios in our imagination that extend beyond what we are strictly given by the fiction). Generally, only *attitude* imagining is considered in the literature on the puzzle of imaginative resistance (see Walton 1990 for an exception). I think that this is a costly mistake; at least some of the confusion in explaining our resistance could be eliminated if we distinguish between the types of imagining that we resist, as in the Sylvan's Box and Oval-Maple Leaf examples. This distinction will also help us evaluate various different responses to the puzzles.

Finally, it is not always clear what is supposed to be the *object* of our resistance. There are two possibilities: the particular evaluative, metaphysical, or conceptual deviance *or* the fictional story as a whole. This is an important distinction. We may be perfectly able to interpret and engage with a story that contains an impossibility *even if* we are unable to imagine a particular impossibility contained within it.

Tamar Gendler (2000) argues that there are two basic ways to explain imaginative resistance. We can be "cantians," "wontians," or some hybrid of the two. *Cantians* about resistance argue that we are often *unable* to imagine certain kinds of impossibilities or evaluative deviances. *Wontians* hold that resistance arises because we are unwilling imagine a situation in which a certain impossibility or deviance is acceptable.

I will examine three accounts of the fictionality and imaginative puzzles. First, Gendler argues that imaginative barriers arise when the principles and background knowledge the reader has accepted in the story leave no way for the impossible or deviant proposition or situation to be true (Gendler 2006). This makes Genlder a cantian about the imaginative puzzle. We are unable to imagine some deviant moral scenarios. However, Gendler is a *wontian* concerning the fictionality puzzle. Even if we could imagine some deviant evaluative response in a fiction, we often *will not allow ourselves* to do so.

Walton takes the opposite approach: whereas Gendler is a cantian about the imaginative puzzle and a wontian about the fictionality one, he argues for the reverse (Walton 2006). I may not imagine a solid gold mountain or a round square, because I have an inability to imagine such a thing. But when I do not imagine a female infanticide, it is not because I am unable to. Rather, it is because I am *unwilling* to. So Walton is a wontian when it comes to the imaginative puzzle (see also Moran 1994). Priest, another wontian, argues that we are able to understand stories that contain inconsistencies like a both occupied and unoccupied box; if we do not imagine them, it is because we are unwilling to.

For the fictionality puzzle, Walton argues that we are *unable* to imagine that some aesthetic or moral evaluation is capable of evoking different kinds of responses than those we are used to. This is because we cannot fully grasp what the world would have to be like for the bad the knock-knock joke to be funny or for female infanticide to be morally acceptable. So Walton is a cantian about the fictionality puzzle; we cannot imagine that evaluative deviances are fictionally true.

There is one more potential solution to the puzzles that I would like to review in some detail: Gregory Currie's notion of *desire-like imagining* (Currie 2002). Like his explanation of the SDP, Curries suggests that positing a distinct attitude can account for our resistance to some

deviant fictional evaluations. Currie suggests that imaginative resistance is the result of an inconsistency in our desires, rather than our moral or aesthetic beliefs. I have not given our desires about fictions any special treatment in this dissertation. In general, I think that we can treat them the way we have beliefs, judgments, and emotions. Like these other states, desires play important functional and inferential roles. Currie suggests that our desires about fictional objects aren't genuine desires they lack those typical roles. So we do not have genuine desires about fictions, but rather *desire-like imaginings*.

Currie explains that we have both internal and external desires about fictional worlds. To use his example, we do not want Desdemona to perish in the fictional world of *Othello*, because we care for her and find her virtuous. This desire is internal, or about, the fictional world. However, we may also possess an "external" desire for Desdemona to die, because her death would make for a better tragedy. Currie contends that these desires are contradictory: we both desire Desdemona's death and desire that she lives. I disagree. On my view, the internal desire is a genuine desire *about the fictional world*. My external desire is about *Othello* as a work of fiction, external to the fictional world. These desires are not contradictory because they possess different types of content.

Currie argues, again for functionalist reasons, that this response cannot adequately explain the functional and inferential roles of desire. Desires motivate action. We would act on our desire if we genuinely wanted Desdemona to survive. We do not act as if we want Desdemona to live, so we lack the relevant desire. This is the standard DAV argument that I have rejected throughout this dissertation. Desires may dispose or motivate us to act, but those dispositions need not lead to actual behaviors. Or, if they do, the behaviors may be different than they would be if the desire was about a real-life person. Such differences can be explained in terms of differences in the

content of our desires.

Let's suppose that Currie is right about desire-like imaginings. How is this supposed to help us solve the puzzle of imaginative resistance? On this view, we lack a genuine desire for Desdemona to live; rather, we have a *desire-like imagining* that she does. The desire runs offline from its normal relations to beliefs and so does not result in behavioral output. Our external desire that Desdemona dies in the fiction is a genuine desire just like any other desire about the real world.

Currie contends that it is relatively easy to suppose that something is true in a fictional world—i.e. to form a belief-like imagining that diverges from things we believe about our actual world. For example, we can easily suppose that a dangerous dragon named Smaug hides in a cave on the Lonely Mountain, sleeping amongst his golden treasure. He even argues that we can easily suppose that female infanticide is morally acceptable in some fictional worlds. However, it is difficult for us to *desire* that female infanticide is acceptable, or even have a desire-like imagining that it is acceptable in a fictional world. Currie argues that we cannot easily construct imaginative replacements of "wicked desires" like we can with beliefs, even if it is possible in some cases. This places Currie in a third camp: he is a *hardian*, rather than a cantian or wontian, about the two puzzles. Deviant desire-like imaginings are not impossible and we may sometimes desire that an immoral act to be fictionally true. However, it is very difficult to get ourselves in the right mindset for this too occur—hence, the phenomenon of resistance. We have a very difficult time imaginatively desiring female infanticide is fictionally true or, or even imaginatively desiring to know what such a world would be like. These desire-like imagining are too far removed from our actual desire that female infanticide shouldn't be practiced.

Currie's view may be able to explain both puzzles with one fell swoop. However, it faces some serious challenges. What is it about desires—as opposed to beliefs—that make them the

object of resistance? We can have all kinds of strange and unethical desires about real-life; why can't we do the same about a world that is not even our own? Currie owes the reader some explanation for why we can easily construct bizarre or immoral *suppositions about fictions*, but not desires and, further, what separates fictional desires from ordinary ones. So Currie's response to the puzzle of imaginative resistance fails.

Gendler's and Walton's explanations also face several problems. First, both suppose that resistance is a robust, commonplace phenomenon. However, my interpretations of the examples in the previous section suggest that resistance may not be as widespread as aestheticians think it is. Second, some authors—including Yablo, Weatherson, and Priest—do not adequately distinguish between what we *imagine* and what we are *willing to accept* as fictionally true. This means that they fail to distinguish between the imaginative and fictionality puzzles. Walton correctly points out that an *inability* to imagine does not equal resistance; we can try very hard to imagine something, but simply be unable to do so (Walton 2006). That does not seem like a case of resisting at all. He suggests that Gendler's cantian response to the imaginative puzzle is mistaken since "resistance" as she uses it conflates *ability* and *resistance*.

Third, some of the traditional responses trade on an equivocation on 'imagination.' Sometimes philosophers seem to be pointing out an inability to *imagistically* imagine a proposition. Priest's box and Yablo's leaf example (also, perhaps, Weatherson's Restaurant at the End of the World) may be guilty of this. However, the real issue in the puzzles concerns *attitudinal* imagining—imagining that something is the case. As Walton suggests, an inability to imagine is not the same as resisting. This is true for imagistic imagining was well. We may be very eager to imagistically imagine a five sided oval, but simply lack the ability to do so.

Fourth, each proposal under-specifies the source of our resistance: is it the fiction in which

the impossibility or normative deviance occurs, or the impossibility/deviance itself (leaving the rest of the story untouched)? Gendler holds the former position; Yablo and Walton suggest the latter. Priest argues for both. This might seem like a trivial point, but I think that it is important for understanding imaginative resistance in general. Just because a story features a proposition or scenario that I do not imagine (either because I cannot or will not) does not necessarily mean that I resist the story as a whole. This is true for any other types of imaginative resistance: contradictions, metaphysical, conceptual, or moral. Consider the sci-fi stories from above, or even *Gone with the Wind*. They all feature instances of imaginative resistance, and yet many people will accept the story *as a whole*, perhaps by "setting aside" the impossibility, like in the Star Trek example. Anyone who makes the stronger claim concerning the fiction as a whole needs to explain why we can or will not accept a fiction if it possesses a point of resistance, especially since we clearly often do.

5. Narrative resistance & immoral learning

On the one hand, it seems like imaginative resistance is an over-blown phenomenon. On the other, there do seem to be some instances in which we resist works of fiction because they present us with deviant moral characters and scenarios. In this section, I will present the second component of my solution to the puzzle of imaginative resistance. Two relatively unexplored reasons why we might seem to resist immoral aspects of fiction are either a) the fiction itself is not constructed in the right way, or b) the reader does not understand a cognitive point that the fiction was intended

to make. Importantly, I resist explaining resistance to fictions in terms of *imagination*. Like Walton, I do not think that there is anything in principle that we cannot attitudinally imagine (see also Camp, in prep). But there are things that we may have *difficulty* supposing to be true in a fiction. The idea is that we can eliminate talk about imaginative resistance and instead discuss our *emotional* resistance, as we'll see in §6. As I've accepted all along, imagination may be involved in our experiences with immoral aspects of a fiction. However, I contend that the so-called puzzle of imaginative resistance isn't a matter of imagination at all.

5.1. Narrative resistance

Some things, like a round square, may be impossible to imagistically imagine (or, at least, are very difficult to imagine). But there does not seem to be anything that we are incapable of *attitudinally* imagining. In my terms, this means that there is nothing in principle that we should not imagine in a fictional world. Most of the time are able to "take up" morally deviant claims and accept that they are true in a fictional world. It may not always be easy to do so, but it is possible.

Consider what would be involved in order for us to accept that female infanticide is morally permissible. We must consider what it would be like for a person to live in certain circumstances and follow certain rules that are very different than our own. We must suppose that it is biologically true that women are inferior to men and the natural, political, and social climate is such that some babies must be killed after birth. It may be very difficult for us to imagine the relevant circumstances, but that does not mean it is impossible.

Gendler and Walton both contend that it is an author's task to get her reader to imagine a story in the right way so that she will accept the morally deviant claim. This is where the aesthetic

and imaginative puzzles collide. An artwork may be aesthetically flawed if it does not enable its audience to be able to think about and consider a deviant evaluative claim. Perhaps the artist has not provided her audience with enough background information about the fictional world or character. Perhaps the morally deviant proposal contradicts an episode that occurred earlier in the narrative. In either case, the audience struggles to go along with the evaluative deviance.

Compare this to my discussion of the SDP. If an artist can get an audience to undergo pro-attitudes towards Satan, Hannibal Lecter, or Humbert Humbert, then why would she not, in principle, be able to get us to imagine that female infanticide is acceptable? Given the right story, we should be able to go along with female infanticide in a fictional world: to think about it and understand what that world is like. A talented artist can get her to audience to engage with even the most morally deviant characters. Like Feagin argues, there may a cognitive payoff for doing so.

I don't think that we can place the blame for an audience's resistance entirely on the artist. Audiences are not limited to what the author explicitly says in the text. Recall my arguments in chapter 5 and 6 that readers each bring their own beliefs and values to a fiction. Our emotional and moral experiences with fiction are a product of our own moral beliefs and emotional tendencies. The same goes with our other beliefs and judgments. We may make judgments and think about a fictional world beyond what the author lays out in a narrative. This can be an important aspect of how we understand and evaluate the work as a whole. So a narrative might not always give us enough information to imagine a morally deviant proposition. However, that does not mean (*contra* Gendler and Walton) that we can't make sense of the deviant proposition on our own. We may have to do extra cognitive work in order to understand a fictional world in which female infanticide is morally permissible.

On my view, then, the imaginative puzzle is actually a matter of narrative resistance. Audiences often find it difficult to accept deviant evaluative claims in a work of fiction either because the narrative lacks sufficient information to them to do so or because they do not properly engage with the story. This makes me a *hardian* about the imaginative puzzle, like Currie.

There remains the question of why audiences might not want to engage with a story that positively portrays deviant moral actions or principles. *Why* would we not want to accept that an immoral action is acceptable in a fiction? What is the source of our unwillingness? This is the basis of the fictionality puzzle.

5.2. Immoral learning

There are very few cases in which we are unwilling to go along with a story and accept a fictional truth. Indeed, if the artist has done her job correctly, we should be willing and able to engage with all sorts of fictional characters and events. But even when she hasn't, the reader may wish to fill in the relevant gaps so as to make the story understandable and relatable. In a sense, the wontians seem to think that some readers/audiences are not playing along with the fiction, just like those viewers of *Star Trek* that cannot look past the metaphysical impossibility of traveling at light speed in order to consider the work as a whole.

I think that the fictionality puzzle—cases in which we *do not* accept moral deviations in the fictional world—is often the result of a worry concerning immoral learning. This occurs when an audience worries that going along with fictional portrayals of immorality in a fiction will cause us to adopt immoral practices and beliefs in our actual lives. Compare the possibility of immoral learning to the fictionality thesis from chapter 6. The fictionality thesis states that we allow ourselves to imaginatively respond to different features of a fiction simply because it is a fiction.

This supposedly frees us to take pleasure in all sorts of things we normally wouldn't in our real lives. However, Gendler suggests that audiences often won't accept fictional truths is because doing so will somehow reflect on one's own beliefs about the *actual* world. We may also worry that there is a causal connection between a fiction and the real-world: reading about a morally deviant character or scenario will cause one to become morally deviant oneself. Plato famously cast the poets out of the ideal state in Book X of *The Republic* for this reason, stating: "their art corrupts the minds of all who hearken to them, save only those whose knowledge of reality provides an antidote" (595d). Fictions often influence our emotions and, in so doing, our moral judgments. Plato worried that audiences might learn bad moral practices from works of art if we are not careful.

It's possible that people sometimes *do* misattribute immoral characteristics to a person based on her emotional and moral responses to portrayals of immorality in a fiction. You might look sideways at a friend who laughs when a fictional character is harmed. However, the fictionality thesis contends that our friend has done nothing wrong here. Her laughing at the fictional character's pain may mean that she accepts that injury is sometimes acceptable in a fictional world. It does not mean that our friend accepts this in the real world. The whole point of fictions is that we recognize that the characters and actions portrayed in a story aren't real.

In fact, I think that refusing to accept deviant fictional truths is sometimes a kind of interpretive mistake. Consider the immoralist position from §2: there may be some aesthetic benefit to the portrayal of immoral actions in a fiction. So if we refuse to accept the deviant fictional truth, then we may also miss out on that aesthetic benefit.

Let's go through an example. Suppose that you and an especially morally-correct friend have just watched an episode of your favorite show, *Mad Men*. This episode, "Mystery Date,"

portrays a great deal of sexual violence against women. The historical backdrop of the story is the (real-life) Richard Speck murders in the late 1960's, in which Speck kidnapped, raped, tortured, and murdered eight nursing students in the Chicago area. Many of the fictional characters in *Mad Men* are equally obsessed with and frightened by the Speck murders (Don's young daughter, Sally, can't sleep after reading about Speck in the paper; her step-grandmother takes perverse pleasure in discussing Speck with her friend on the phone). Other characters simply carry about their business. The creative ad-man, Don Draper, and the young copy writer, Michael Ginsberg, promote a sales campaign for Butler footwear. Ginsberg's final pitch highlights the backdrop of sexual violence in the episode:

> 'We were gonna come in here and talk about Cinderella, but it's too dark…I mean she's running down this dark side street, and it's outside a castle so it's got those walls and the cobblestones. And she's running but she's only got this one incredible shoe for her incredible gown, so she's hobbling—wounded prey. She can hear him behind her, his measured footsteps catching up. She turns a corner, those big shadows. And she's scared. And then she feels a hand on her shoulder and she turns around, and it doesn't matter what he looks like. He's handsome at that moment, offering her her shoe. She takes it. She knows she's not safe, but she doesn't care. I guess we know in the end that she wants to be caught' (*Mad Men* episode 5.4, "Mystery Date").

The Butler people gobble it up, oblivious to the menacing tone of this scene. Later in the episode, Don has a dream in which he strangles his mistress after they sleep together. Finally, the episode raises the specter of the office manager Joan's rape by her then fiancé, Greg.

Some people may take delight in "Mystery Date." The whole episode is dark and gloomy, much more muted in both color and tone than a typical *Mad Men* episode. Maybe you take pleasure in hearing about the Speck murders, in watching Don's bizarre dream, and Ginsberg's haunting Cinderella tale. Your morally correct friend might be deeply disturbed when she realizes that you take delight in a show that nonjudgmentally portrays violence against women. In fact, it's unclear

whether the show condones or condemns Don's immoral behavior (how he treats women, his wife, his children, his employees, etc.). Ginsberg's sales pitch is mainly used to contrast the personalities of the dour, stern Don and the impulsive young copy writer. Only clear judgment is against Joan' husband, Greg, who so clearly disrespects his wife's desires and ends up alone in Vietnam. The moral ambiguity with which the narrative presents Don and the others bothers your friend, who resists any glorification or even casual acceptance of the show's portrayals of immorality.

Most people don't think that violence against women is ever acceptable in the real world. But *Mad Men*'s fictional world is just that: fictional. Don and Joan's world may be very similar to the actual world, but no real person suffers as the result of the events in this episode. Accepting the sexual violence against the women in the show does not mean that one accepts violence against women in the real world. You accept that this may be *fictionally* true, but not *actually* true. No one is harmed by your taking delight in the portrayals of immorality in "Mystery Date."

Fictions often portray immoral characters or actions that are not condemned in the narrative. Like Kieran and Feagin, I argue that, generally speaking, these portrayals of accepted immorality are supposed to prove some cognitive point. The audience is supposed to take the portrayal of immorality as an opportunity for moral reflection and learning. You may watch "Mystery Date" and enjoy all of its dark, immoral scenes. Maybe you don't think about the significance of the sexual violence at the time you watch it. You simply enjoy the episode. Upon reflection, though, you may realize that there might be a deeper significance to some of the violence in the episode. *Mad Men* is a partly a satire that reflects the social mores of both the 1960's and the present. The ad-men casually accept violence against women and using women as mere sexual objects. The show holds up the image of sexual violence in the show to reflect our own society, either as it was or currently is. So watching portrayals of immorality on screen should

cause the viewer to reflect on her own moral beliefs.

This, perhaps, is the cognitive point that "Mystery Date" and other *Mad Men* episodes attempt to make. Immoralists would count this opportunity for moral reflection as an aesthetic virtue, something that makes *Mad Men* a good show. Indeed, we could contrast "Mystery Date" with a show that simply glorifies sexual violence against women without making a similar cognitive point. It may be right to resist the condoned violence in this case.

One could protest that it may be quite difficult to tell when a fiction attempts to make a point by portraying violence and when it's simply portraying violence for shock value. This is true; it's up the viewer to be able to draw out the cognitive point in some narratives and to decide for herself when portrayals of immorality are unacceptable. Here, too, the worry is that taking pleasure in fictional immorality will somehow transcend into real world immoral beliefs or actions. I will return to this point in the following chapter when I discuss moral learning from fictions.

The *Mad Men* example I give here is intended to show that the fictionality puzzle may sometimes be a kind of interpretive mistake. Audiences are typically quite willing to engage in a morally deviant fiction so long as the narrative is presented to us in the right way, with a sufficiently detailed plot, fictional world, the right cast of characters and, most importantly, an interesting or important cognitive payoff. I do not think that it is a coincidence, or even a matter of logistics and practicality, that the morally deviant propositions in the articles on imaginative resistance are made-up cases by philosophers, not actual or complete fictions in themselves. If we take a work of fiction as a whole—as Feagin says, a *temporally extended object*—then my guess is that we will rarely resist accepting that some morally deviant proposition is true *in the fiction*. Importantly, this does not mean that we need to consider the morally deviant proposition as true *in our actual world*—even if the fiction presents us with a world that is remarkably similar to our

own.

6. Emotional resistance

I have tried to explain away some cases of supposed imaginative resistance. Surely, though, there are *some* genuine instances of the phenomenon. Genuine imaginative resistance amounts to an unwillingness to imagine deviant evaluative claims in a fictional narrative. In this section, I will argue that these supposed cases of imaginative resistance are really instances of *emotional* resistance. I will focus on deviant moral claims, but my arguments apply to aesthetic evaluations as well. Our resistance to positive portrayals of immoral characters and scenarios are due to a difficulty in overcoming negative emotional responses to them.

Emotional resistance explains why we might only resist some immoral features of a fiction. This is a challenge for other explanations of imaginative resistance. None of them explain what it is about *certain* examples of moral deviances that cause us to resist them. As we saw with the SDP, there are many instances in which we respond positively towards immoral characters. We either overcome an initial resistance to the character or we experience no resistance to her at all.

In fact, the fascinated attention account of the SDP helps explain the puzzle of imaginative resistance as well. A narrative can be constructed in such a way that we will feel pro-attitudes towards an immoral character. We are willing to accept deviant characters and actions if the narrative is presented in the right way. We may emotionally resist immoral characters or scenarios in a fiction if we are not sufficiently fascinated by a character or if we are not shown his or her sympathetic features.

Moreover, a multi-level appraisal theory of moral judgments perfectly explains our resistance to immoral aspects of fictions. Resistance may be the result of fast, subpersonal negative reactions to portrayals of immorality. Suppose that you are watching *The Godfather*, a film that not only presents murder, organized crime, drug trafficking, kidnapping, and theft, but often glorifies these actions. Do we resist *The Godfather* or not, and why? If so, is the resistance to the story as a whole, or an individual act within the story?

At the end of the film, Michael Corleone has the Dons of the other major crime families sent "to sleep with the fishes," thereby assuring his own dominance in New York mob scene. You watch as one man after another is killed, contrasting with the sanctimonious shots of the baptism of Michael's nephew. What is your initial emotional response to this scene? Some people might feel a sort of malicious glee due to their sympathies for the Corleone family. Others might experience disgust, horror, or revulsion at the murders. According to the multi-level appraisal theory, one's disgust, horror, or revulsion is the result of the initial affective appraisal of the murders that may or may not be consciously experienced. A positive response might suggest that the viewer accepts Michael's actions, at least initially.

From here, slower, cognitive appraisals of the murders might come about. The cognitive appraisal may overturn the initial affective appraisal. We automatically have a negative response to Michael's actions but later decide that his was in the right. We do not resist his immoral deeds. Alternatively, our cognitive appraisal of the murders might *not* overturn the initial negative affective appraisal. Perhaps the viewer is sickened by the depravity of all the crime families. These murders mark Michael's moral turning point and herald in a new era of morally bankrupt, avaricious organized crime. We resist the positive portrayal of the Don's murders because we cannot overcome our initial negative reaction to them.

It is difficult to overcome our initial affective responses towards *any* situation: fictional or non-fictional, moral or amoral. Imagine a scenario in which you suspect that your partner has cheated on you. You notice all the late nights at work, the mysterious phone calls, and the unaccounted for afternoons. You experience intense jealousy, rage, and hurt. But then you discover that your partner was *not* cheating, but planning your surprise birthday party. You may instantly feel relief while at the same time experience some residual hurt or jealous affective feelings. It's challenging to overcome our affective appraisals and feelings. We may have to actively work towards overcoming them. This could also apply to our emotional responses to fictions; we may have to remind ourselves of the point of the story and take other measures to eradicate our initial negative response towards the murders (see Harris 2000).

Finally, it is possible that the viewer has misinterpreted the story. It's not entirely clear that *The Godfather* as a whole *does* condone the assassinations. Consider how hard Michael worked to prevent himself from becoming involved in the family business, the tragedy of Apollonia's death, Connie's dismay after Michael has her husband killed, or Kay's horrified look at the end of when Michael receives his *caporegime.* Perhaps, then, our viewer does not resist the positive portrayal of murder because there is no positive portrayal to begin with!

As I've argued, we can make a similar point can be made for many stories that portray immoral acts: they do so for some cognitive point, not because one is actually supposed to accept them. We may have difficulty in accepting a character or action in a fiction because we typically associate strong negative emotions towards such characters and actions in real life. It may be that we are not given enough information about a character to care about her or to understand how an immoral action—like Walton's female infanticide example—could possibly be true. Thus, our initial affective appraisal of the character or scene would not be modified by later cognitive re-

appraisals. This could lead to our resistance: we refuse to watch a film that positively portrays immoral actions or we cannot accept that the immoral action is acceptable partly because we have no reason to update our initial negative reaction.

7. Dissolving the puzzle

My goal in this chapter was to show that the puzzle of imaginative resistance is not really all that puzzling. Standard cases of imaginative resistance do not adequately motivate a robust puzzle to be solved. They are not genuine examples of resistance or involve imagistic rather than attitudinal imagining. Imaginative resistance is not as prevalent a phenomenon as is typically supposed. When resistance does occur, it may often be due to a narrative or interpretive failure: what I have called narrative resistance and the challenge of immoral learning, respectively. Finally, I have attempted to show that imaginative resistance is better understood as *emotional* resistance—or, at least, emotional *difficulty*.

If imaginative resistance is a genuine worry, it does not concern fiction alone (see also Matravers 2014). We often encounter real-life stories that we resist due to moral deviances, such as *Triumph of the Will*. We resist these stories for roughly the same reasons we would resist any morally unacceptable action: we disagree with it. As in the purely fictional cases, we may resist accepting the moral deviances in these works. This may be a matter of how difficult it would be to change one's moral outlook in order to accept the positive portrayal of Nazism. It may also be a matter of overcoming our strong negative emotions towards Nazism, which would be quite

difficult to do.

Chapter 9: (Im)moral Learning from Fictions

1. Learning from Dexter Morgan

The last puzzle that I will discuss concerns the possibility of learning about real-world morality from fiction. This is an extension of my arguments from the previous chapter: one reason why the challenge of immoral learning is so persistent stems from the worry that portrayals of fictional immorality will carry over into real world actions and beliefs.

Take the TV show *Dexter*. This show portrays the escapades of Dexter Morgan, Miami Metro blood-spatter analyst by day, self-described psychopathic serial killer by night. Audiences typically want Dexter's plots to succeed, despite his murderous tendencies and sometimes callous demeanor. We do not want him to be caught by the police even if his capture would be morally justified all things considered. This is partly because Dexter is a serial killer with a code of honor. He only kills people who (he thinks) deserve it: drug dealers, other serial killers, rapists, mobsters, etc. Dexter follows this code to the tee and rarely strays from it. In fact, Dexter blames other serial killers for not following his code and berates himself when he slips and impulsively harms an innocent.

Another reason for our sympathy for Dexter is that he is not, strictly speaking, a complete psychopath. He occasionally feels remorse and he often reflects on the morality of his actions. "How evil could I be?" Dexter muses, as he runs off to help out a neighbor with some handy-work. Dexter fits my fascination hypothesis from chapter 7: he is intelligent, attractive, loyal, out of the ordinary, and interesting.

I do not think that *Dexter*'s viewers will take this code to heart and take up murdering neighborhood criminals. We discover the cost of Dexter's type of vigilante killing—for instance, the death of Dexter's wife and his isolation at the conclusion of the series. But there may be a sense in which we do learn *something* from Dexter's immoral behavior. While some characters, such as his sister Debra and his friend Lumen, start out repulsed by Dexter's behavior, they come to admire it. Dexter is never caught or punished by the authorities. He is portrayed as a generally nice guy with a bit of a killing problem, and certainly a better moral agent than many of the people he murders. Perhaps the worry is that accepting a serial killer—no matter his motives—will cause us to eventually accept similar people in real life. Feelings about violence and vigilante crime in the fictional world of the TV series could bleed into a viewer's thoughts and actions about the real world. Call this *immoral learning*.

I introduced Jenefer Robinson's notion of a "sentimental education" from fiction in chapter 5. She argues that fictions allow a reader to gain experience in new emotional territory and gain self-knowledge about emotional tendencies that we already possess. If we can learn about our emotions from fictions, then it seems likely that we can learn about morality from fictions as well, especially if morality is based on emotional responses.

In this chapter, I will explore the possibility of moral learning from fictions. I will begin by discussing the plausibility of learning from fictions in general. I will examine several arguments both in favor of and against this possibility. In §3, I will connect the moral learning debate with another hot debate in the philosophy of film: the question of whether film can *do* philosophy. I offer some evidence in support of the idea that fictions can do positive philosophy, as thought experiments. I contend that the moral learning and film-as-philosophy debates are closely related, even though they have not been understood that way in the aesthetics literature. In fact, the idea

that some films can do philosophy is evidence that they can teach us about morality. I will end this chapter by reexamining the possibility of *im*moral learning.

2. Learning from fictions: Nussbaum vs. Currie

2.1. The optimists

Robinson claims that readers can gain a sentimental education from fictions. Fictions engage our emotional systems and, if we reflect on those emotions, we may acquire self-knowledge about our own values and desires. We can also practice mindreading of fictional characters and consider various hypothetical alternatives to emotional situations. Other philosophers have argued that we can gain a *moral* education from engaging with fictions, and literary works in particular. Susan Feagin (1983) and Martha Nussbaum (1990) take their cues from Aristotle (1947), arguing that engaging with fiction plays an instrumental role in the development of one's moral and intellectual character (see also Gendler & Kovakovich, 2005).

Nussbaum claims that there are two ways in which fictions can aid moral learning. First, fictions often present us with interesting and original moral content. She provides an extended example of the character Maggie from Henry James's novel, *The Golden Bowl*. Nussbaum explores how James presents Maggie's idiosyncratic, perfectionist conception of goodness and fear of doing harm. The narrative suggests that Maggie's absolute convictions are overly idealistic and simplistic. The novel-length treatment of Maggie's personal relationships sets up several moral

problems and then "solves" those problems by determining each character's fate (Nussbaum 1990, 132).

Of course, we could simply read about morality in a philosophy text. However, Nussbaum doubts that readers would learn as much about the particular moral question at hand from a philosophy paper as she would from James' novel. Sometimes readers need particular cases and examples in order to understand the complex social and practical implications of a moral principle. Short, abstract philosophical works often cannot do justice to the intricacies of moral dilemmas in our actual lives.

Fictions provide us with unique moral content to expand our understanding of moral values and principles. Engaging with fictions also develops our moral skills. Presenting moral dilemmas and principles in a story calls upon our moral faculties: our attention to morally relevant details of a situation or person, reflection on our own moral behaviors and how others act in moral situations, and exploring and weighing options in order to make moral decisions. Fictional dilemmas may also trigger our moral emotions for a kind of "knowledge by acquaintance," what I called *experiential knowledge* in chapter 4. Fictions grant us access to situations and personality types that we would never encounter in our daily lives. We practice our moral and emotional responses on these fictional entities and learn from example. We can then apply any know moral insights we gain to our own lives.

Some philosophers have argued that fictions can provide us with more general types of learning, besides a sentimental or moral education. Peter Lamarque (2009) lists five different categories of learning from fiction: vision, imagination, learning what it is like, conceptual knowledge, and cognitive strengthening. Note that these learning categories were originally thought of in terms of literary fiction, but I will make the case that they can apply to any fictional

medium, at least in principle

First, philosophers such as Iris Murdoch (1970) and Hilary Putnam (1978) have argued that fictions educate readers by helping them picture and understand real-life situations. The *vision* of a work should be understood metaphorically; readers learn to view the fictional world in a certain way, and apply that understanding to the real world. We learn an author or narrator's vision perspective on the world. As Putnam (commenting on Celine's *Journey to the End of the Night*) remarks: "I do not learn that love does not exist, that all human beings are hateful and hating…What I learn is to see the world as it looks to someone who is sure that hypothesis is correct" (*ibid*, 89; quoted in Lamarque 2009, 241). Perhaps I could previously imagine what it would be like to learn that love doesn't exist (for example), but encountering this perspective in a fiction adds a new layer of experiential knowledge that I didn't formerly possess.

Second, readers actively engage their imaginative faculties while reading works of literature. In chapter 1, I discussed van Leeuwen's three of types of imagining: imagistic, attitudinal, and constructive. While reading a novel, we may imagistically call to mind characters, constructively imagine scenarios and propositions true of the fiction, and attitudinally imagine fictional objects. This kind of imaginative engagement with fictions may broaden the scope and power of one's imaginative abilities in other areas, such as hypothetical or forward-looking thought. This is true even if, as I have argued, imaginative engagement is not necessary for fictional experiences.

Third, fictions may teach us what it is like to be someone with different characteristics than the reader possesses, or to be in a unique or unfamiliar scenario. Nussbaum and Robinson both argue for this position. This is another kind of experiential learning; we may learn what it is like to be a certain kind of person or in a unique moral situation. For example, I could begin to learn

what it is like to suffer from severe depression after reading *The Bell Jar*. I may not learn exactly what such a deep depression feels like, but I would have some idea of its effects on one's daily life.

Fourth, fictions may teach us about folk psychological concepts. We can gain self-knowledge of emotional tendencies, desires, or values that we did not know we possessed (Nussbaum 1995, Robinson 2005). It may even be possible that fictions can grant us propositional knowledge about folk psychology, a possibility I will explore in greater detail momentarily.

Finally, one might contend that fictions are unlikely to teach us anything new about morality or emotions (for instance), but they may nevertheless enhance knowledge that we already possess (Gibson 2003). We gain moral skills from fictions much in the way we would by practicing sports: through the repeated application, development and enhancement of skills.

In sum, we can potentially gain three types of knowledge from fiction: propositional knowledge about psychological or social concepts, knowledge of how to apply those concepts and moral/emotional/social skills, and experiential knowledge of what it is like to be a particular type of person or in an unfamiliar situation. In what follows, I will make the case for all three types of *moral* knowledge. Each of the previous styles of learning can be applied to a reader's moral education; we can gain moral perspectives, learn about moral concepts, and practice moral skills while engaging with some fictions. We may even gain self-knowledge about our own moral responses to unfamiliar people and situations. Indeed, it seems like a platitude that engaging with great works of fiction makes us more morally sensitive. What reasons do we have to doubt this claim?

2.2. The skeptics

It may seem like the authors in the previous section all assume the claim that we can learn from fictions without putting forth any real argument in its favor. But what evidence do we have to support this claim? Currie (2013) notes that philosophers like Nussbaum, Lamarque, Robinson, etc. all assume that we learn psychological and moral truths from fictions. Unfortunately, these philosophers do not provide any empirical evidence to support their optimism about moral learning. Currie claims that there is such no empirical evidence. It may *seem* like reading great works of fiction—those we would consider to be psychologically intuitive and revealing, such as novels by Dostoevsky, James, Tolstoy, and Austen, or films by Godard, Mehlville, or Antonioni—make their audiences more insightful, thoughtful, reflective, and morally and emotionally sensitive in the course of their daily lives. But it is equally possible that readers who are already reflective, insightful, thoughtful, and morally and emotionally sensitive are naturally drawn to great works of literature that engage these faculties.

The optimist might protest that recent studies show that we *can* learn from fiction. In one, the psychologists David Comer Kidd and Emanuele Castano (2013) found that people who read literary fictions as opposed to or pop fiction do better on empathy and mindreading tests. Engaging with literary fictions (such as a work by Anton Chekov) primes readers for empathetic thoughts and behaviors. This conclusion seems to support the Robinson/Nussbaum line that we can gain a sentimental and moral education from literary fictions, especially if we think that social cognition and empathy are important components of how we morally engage with others.

The skeptics have a ready reply to these results. As the researchers note, it is unclear how long these effects last—for a few hours? Days? Minutes? It's not clear that we should count this

as moral learning if the participants were merely *primed* to respond empathetically. It is also possible that the differences in responses to empathetic and mindreading tests stem from how the narrative was constructed and what faculties it employs and not necessarily from the fictional content (what it supposedly reveals about the human psyche, for example). One final interpretation could be that people who tend to be better at mindreading and empathy tests also happen to be frequent or talented consumers of literary fiction. Reading the literary work draws upon a skill that they already have.

The skeptics have another reason to doubt that we can gain knowledge from fictions. Even if we can learn about morality or psychology from fictions *in principle*, most authors rely on faulty psychological theories of character, intention, and decision-making. If this is so, then the great works of fiction would not help us learn about how humans think (Currie, in prep). Currie sites recent research in cognitive psychology on character and decisionmaking that undermines how standard fictional characters think and behave.

First, some contemporary philosophers and psychologists have called into question the notion of *character* as it is employed by virtue theorists (see Doris 1998 & 2002). Virtue theorists seem to be committed to the existence of persisting character traits that help explain how people will act: honesty, courage, loyalty, generosity, etc. An honest person will refuse to lie to their friend or to a stranger and will also refuse to cheat on a test or steal from a store. It is assumed that a person's character trait will be reliable across various situations and over time (see Sreenivasan 2002).

Against this notion of character, studies like Stanley Milgram's electric shock experiments and Philip Zimbardo's Stanford Prison Experiment seem to suggest that ordinarily virtuous people—who might appear to possess certain basic, virtuous character traits—can be made to do

awful things when pressured to do so. *Psychological situationists* argue that our behaviors may vary dramatically in different situations and do not raw upon constant personality traits. In John Darley and Daniel Batson's "Good Samaritan" study (1973), Princeton theological students (ideally the utmost empathic and caring individuals) were studied to see if they would come to the aid of a needy bystander. They found that self-ascribed traits such as empathy and compassion were not suggestive of who would help the bystander. Rather, the determinant was which students *were in a hurry*. Students who were told that they was running late for a lecture were significantly less likely to help the bystander than if they were early or on time.

Based on research like this, *philosophical situationists* have argued that the traditional virtue theoretic notion of a constant moral character is misleading (Doris 1998 & 2002, Harmon 2000). Kevin Timpe (2008) describes philosophical situationism in terms of three related claims:

> 1. *Non-robustness*: moral character traits are not robust—that is, they are not consistent across a wide spectrum of trait-relevant situations. Whatever moral character traits an individual has are situation-specific.
>
> 2. *Consistency*: while a person's moral character traits are relatively stable over time, this should be understood as consistency of situation specific traits, rather than robust traits.
>
> 3. *Fragmentation*: There may be considerable disunity in a person's moral character among her situation-specific character traits.

If these claims are true, then situationism seems to present a problem for works of fiction. Currie argues that great works of fiction present characters that possess constant character traits that shape their behaviors. Examples abound: Luke Skywalker, Harry Potter, Frodo Baggins, Ned Stark, Jane Eyre, even Batman all seem to act on the basis of constant character traits that are not easily swayed by situational features. Rather, the character brings these traits to bear on difficult moral situations.

Jane Eyre is honest, so she consistently acts honestly across most situations, even when she has the perfect opportunity to act dishonestly. Furthermore, Jane possesses a unity of virtues; she is not only honest, but loyal, brave, and hardworking. Her virtues are not fragmented. Situationists would argue that his portrayal of a young woman is unrealistic and does not capture how people actually behave or the values they actually possess.

Currie also argues that fictions present a faulty picture of decision making and causal reasoning. For example, a narrative may present a character as faced with a moral dilemma and she must reason through various options in order to reach a satisfactory conclusion. Her reasoning serves as the main causal basis for the choice she makes. Why does Jane Eyre decide to leave Mr. Rochester, the man she loves, in favor of an uncertain fate? Because she cannot bear the idea of throwing aside her moral beliefs for a life of dishonesty and callousness to others. Jane seemed to have conscious access to all of her reasons for leaving and was able to justify her actions to others later in the story.

Recent work in cognitive psychology has brought this picture of decision-making into question. In a famous study on the introspection illusion, Richard Nisbett and Timothy Wilson (1977) presented participants with two pairs of stockings. The participants were able to present justifications of their decision when asked which pair they preferred. For instance, they liked the color better of one of the pairs; one of them looked like it was of a higher quality, etc. In fact, both pairs of stockings were the same! The participants tended to prefer the pair on the right-hand side for no other reason than that it was on that side. The verbal justifications for their decisions were merely confabulations.

Similarly, Daniel Wegner (2002) has argued that conscious will is an illusion. We think that our conscious thoughts cause our actions, when, in fact, it is possible that our brain primes

action before conscious thinking takes place. We do not know that brain states give rise to actions, of course, because these processes are conducted unconsciously. So we conclude that conscious thoughts must have caused our actions. Benjamin Libet (1985) makes similar claims in his studies on the timing of conscious voluntary actions (see also Graves et al 2010, Haggard 2005, Haggard & Libet 2001).

Currie concludes that evidence from cognitive psychology on situationism, decision-making, and the conscious will undermines the idea that we can learn about psychology and morality from fictions. Authors present fictional characters as possessing predictive character traits, as having access to their own intentions and deliberative decision-making processes, and as consciously willing to act. However, if the psychological data is correct, people don't really think or act the way fictional characters do. We are mistaken in thinking that we can learn about such things from fiction

Perhaps we can gain *pretend* knowledge from fictions, as Currie suggests. We can approach a literary work *as if* it presents a true picture of human psychology and see what we can conclude on that basis (Currie, *in prep*). This is a natural implication of the DAV; we do not have stereotypical mental states about fictions, so how can we learn anything about real-life mental processes from them? Coupled with the skeptical evidence from cognitive psychology, it does seem like the best we can learn from fictions is how we think mental states *might* work, not how they *actually* do.

3. Withstanding the skeptics

I want to resist Currie's skeptical arguments. The optimist position faces two challenges. First, we must respond to Currie's claims concerning character and decision making in order to show that it is possible for audiences to learn about fiction. We need to determine whether authors present us with accurate depictions of the human psyche. But even if we can eliminate this negative claim, the optimist still lacks a positive account of how consuming fictions makes us better moral agents. I will turn to this challenge in the next section.

We can question Currie's quick acceptance of the situationist literature as evidence against the psychological reality of character. Currie's situationist argument that character traits are weak and unstable has not been universally accepted in the psychology and philosophy literature on virtue (Annas 2005, Sabini and Silver 2005, Sreenivasan 2002 & 2009, etc.). Many virtue ethicists argue that character traits need not entail *inflexible* or *scripted* behaviors. Rather, they are *predictive*. It may be that a character trait for honesty disposes a person to tell the truth, and she will in most circumstances. However, character traits may be flexible and context sensitive. An honest person will, *ceteris paribus*, act honestly. This does not mean that she *always* will. An honest response can be defeated by context sensitive factors and knowledge (Sreenivasan 2002).

This means that fictions may actually present us with an acceptable picture of moral character. Do the brave characters always act bravely and the dishonest characters always act poorly? Not necessarily. Inflexible fictional characters are boring and unrealistic. We are fascinated by a villain with a compassionate streak and a generally honest protagonist who sometimes lies or cheats. We like our fictional characters to be somewhat unpredictable. There are also characters that seem to have stable personality traits who nevertheless change their minds due

to the context in which they find themselves (e.g., Huck Finn and Anna Karenina).

What about Currie's argument about conscious decision making? Do characters in the great works of fiction always act on the basis conscious reasoning? This research presents a greater challenge for the optimist. Decision making and deliberation does seem to be generally unconscious and subject to unknown or inaccessible influences and biases. In contrast, fictional characters have access to their own mental processes and states.

However, there are some literary works that show their characters acting on the basis of emotional feelings and intuitions, or simply not knowing why they acted as they did. Dostoevsky's *Crime and Punishment* is my favorite example of this. Raskolnikov decides to go through with murder of the old pawnbroker and her sister on a whim, even though he considered it for a long time. He spends the rest of the novel agonizing over his motives (which vary depending on his mood). Consider also the stream-of-consciousness techniques in works by the British modernists. Virginia Woolf's *To the Lighthouse* and James Joyce's *Finnegan's Wake* go to great lengths to show that people do not act on the basis of orderly, deductive causal chains, but rather from general impressions, moods, and memories. Of course, one could still argue that the stream-of-consciousness is supposed to be conscious and so does not present a counterexample to Currie's claim. But it does show that our actions do not always proceed from rational deliberation.

Even if *most* literary works present us with inaccurate pictures of conscious decision-making and deliberation, not all of them do. It is possible, then, that fictions can provide us with some knowledge about how we real people think and make decisions.

Moreover, it is far from obvious that artists are really concerned with presenting scientifically accurate portrayals of psychological processes, or even that they take their stories as a source of learning (moral or otherwise). Artists may be unconcerned with presenting a picture of

rational thought that squares with our best cognitive psychological and neuroscientific theories. Instead, they write about how most people *appear* to act and make decisions. Artists present characters in terms of what appear to be, *prima facie*, plausible psychological processes. Artists offer us the building blocks of folk psychology. So the value of this approach may depend on the value of folk psychology in general.

Imagine a fictional story with an omniscient narrator that does not explain why the protagonist behaves as he does. The narrative would be quite confusing and chances are readers wouldn't find it very engaging. The reader would have to do a lot of work to understand how the character thinks and feels. There may be some challenging, scientifically accurate stories out there. It may be purposively difficult to understand the narrative of these stories, since everything we know about a character typically comes from what we are told about her in the text. We don't have real world context or perceptual cues to inform our understanding of the character. Perhaps such stories make some cognitive point about the incomprehensibility or opacity of human behavior. It would be an interesting philosophical and psychological case study, but maybe not the most engaging story.

We tend to treat fictional characters the way that we do real people. We assume that their actions are motivated by character traits and reasons. This is part of what it means to take the fictional stance; we recognize the character as a person like one in the real world, and doing so helps us to understand the character. It could be that folk psychology is wrong in general. So we may be mistaken about the decisions and virtues of fictional characters—but, then, we are also wrong about real people. This is interesting and noteworthy for those working in cognitive science, but knowledge of situationism and the illusion of conscious reasoning will likely not change how we treat either real people or fictional characters.

4. Fictional thought experiments

I have been fended off Currie's charge that fictions utilize inaccurate notions from folk psychology. There is another hurdle to cross before we can say that we can learn about morality from fiction. The claim that we can gain *any* kind of knowledge from fiction is susceptible to Currie's challenge that there is no real empirical evidence in favor of the view that consuming fiction enhances our psychological understanding.

In this section, I will make the case that some fictions are *extended thought experiments*. If we think that we can learn about morality from thought experiments—which philosophers and cognitive scientists take for granted—then there is no reason why, in principle, we should deny that we can also learn about moral issues from some works of *fictions*.

There is a growing literature on the relationship between philosophy and film. Some philosophers argue that film can *do* philosophy. Others deny this possibility. My arguments will draw from this literature, but will not be limited to film fictions. Likewise, Currie's claims concerning literary fictions carry over to films. Film characters also seem to possess steady personality traits and use conscious deliberation to make decisions that guide their behavior. Indeed, many of the challenges facing film as a form of philosophy apply to literature and other fictional media. I will point out differences between art forms when appropriate.

I will begin with several arguments against the claim that film can do philosophy. Like many authors in this area, Bruce Russell (2000) grants that films can be counterexamples to philosophical points or address philosophical theses. Woody Allen's *Crimes and Misdemeanors*, for example, presents a counterexample to Plato's argument in *The Republic* that only virtuous people can be truly happy. *The Matrix* illustrates Cartesian external world skepticism. But these

films do not actually *do* philosophy. While they illustrate or present counterexamples to philosophical claims. But neither one presents original, unique, positive philosophical ideas.

Right away we can think of one good response to Russell's skepticism. First, why is it that counterexamples are not "real" philosophy? Thomas Wartenburg (2007) reminds us of Gettier's short paper "Is knowledge justified true belief?" (Gettier 1963). This paper presents several brief counterexamples to the claim that justified true beliefs are necessary and sufficient for knowledge. No one would deny that Gettier's paper *is philosophy* even though it does not present a positive theory of knowledge. Wartenburg (2006) also points out that many of the philosophy papers published in journals each month do not make unique contributions to their area. Instead, they may elucidate, bolster, or reject established philosophical claims. Russell's notion of what it means to "do philosophy" is too limiting.

Russell also points out that films cannot establish truths about the actual world. He states that: "[an] imaginary situation cannot supply real data" (Russell 2000, 390). Films present us with specific moral examples, but we cannot learn about real-life from them. This is because the fictional examples cannot be generalized or the medium is incompatible with doing philosophy. This charge proves more difficult to argue against without some positive account of how films can do philosophy, so I will hold off on responding for the time being.

Another skeptic, Paisley Livingston (2006), also claims that film has limited philosophic potential. Livingston's arguments form two horns of a dilemma. Underlying each is the notion of *medium specificity* (see Carroll 2008). In order for film to do philosophy, one would have to show how the film medium can uniquely capture philosophic ideas in terms of its unique qualities. These features of the film medium that are unique to it, such as its capacity to depict movement. The first horn of the dilemma accepts this medium specificity claim, but notes that "exclusively cinematic

insight cannot be paraphrased" in the way that a philosophical argument requires (*ibid*, 12). The medium specific qualities of a film—its ability to utilize editing, camera movement and angle, focus, correlations between sounds, music, and a shot, etc.—cannot capture succinct, verbally articulated arguments as philosophy does. This is one area in which literature may have an advantage over film—literature is linguistic, so there is no medium specific reason why literature cannot do philosophy.

The second horn of the dilemma rejects the medium specificity requirement. However, doing so amounts to a trivialization of the notion that film can do philosophy. An important feature of film is its ability to visually and aurally represent and record ideas and stories. Livingston argues that these general cinematic capacities merely *represents* philosophy—for instance, presenting a character as spouting philosophical theses. The film *itself* cannot make a unique philosophic contribution in virtue of its medium specific capacities to visually and aurally represent and record ideas. Thus, accepting either horn of the dilemma bars film from making a unique contribution to philosophy. Like Russell, Livingston holds that films like *The Seventh Seal* can be a useful complements to philosophy, especially for pedagogical purposes in illustrating philosophical points. However, Livingston denies that these purposes amount to positive philosophy.

In sum, there are three main charges against film as philosophy (Wartenburg 2007). First, there is the *explicitness objection*: film as a medium cannot capture the conceptual determinacy of philosophy (1st horn of Livingston's argument). Second, the *generality objection* states that philosophy is concerned with general concepts about reality, whereas film can only capture particular examples that do not generalize (Russell's objection). Finally, the *imposition objection* states that film can be used for philosophically interesting purposes, like providing counterarguments or illustrating philosophical ideas, but cannot add to positive philosophy (both

Russell and the 2nd horn of Livingston's argument).

Russell and Livingston make a universal claim against the possibility of film as philosophy: no films can do philosophy due to the nature of the film medium. In general, optimistic accounts of film and philosophy carefully avoid universal statements. It's not that *all* films can do philosophy, but rather that *some* can. For instance, Carroll argues that there is at least one case of film as philosophy: the short structuralist film, "Serene Velocity" (Carroll 2006). Carroll argues that "Serene Velocity" presents an argument of sorts concerning the nature of the motion picture, similar to Andy Warhol's film *Empire*. Structuralist films are version of *minimalism*. Minimalist visual artists like Frank Stella, Sol LeWitt, Donald Judd, and Robert Morris created works that attempted to capture the essence of their art form. They stripped away what they deemed to be all the unnecessary components of painting, such as the representations of objects. Doing so reveals the true essence of painting: color, form, shape, or line. Carroll argues that the goal of "Serene Velocity" is similar to those of the minimalist painters. It is a kind of "metafilm"; a film that comments on and puts forth an idea concerning the real essence of cinematic art. The way the narrow hallway is still, yet also seems to move with the changing camera zooms creates a sense of the possibility of movement which, indeed, is the essence of cinematic art (*ibid* 178; see also Danto 1979).

If Carroll is right, then we have at least one example of a film that does positive philosophy even to Livingston and Russell's standards. Of course, this example might not persuade us that film in general can do philosophy. After all, it does not present a general philosophic idea, such as free will or mind-body dualism. We still need a way to respond to Livingston and Russell's point that film can present us with general philosophic ideas in areas like metaphysics, epistemology, and ethics.

I think that there is an answer to this charge. Both Carroll and Wartenberg suggest that films can also do philosophy by acting as thought experiments. One might think that thought experiments can illustrate philosophical points or even draw out philosophical conclusions, but they are not actual works of philosophy. As we have seen, philosophers like Russell and Livingston deny that illustrative thought experiments can do positive philosophy. But it is unclear why this is so. After all, many philosophers utilize thought experiments in the course of their arguments. Plato's Allegory of the Cave and the Myth of Gyges from *The Republic* are two such cases. Thought experiments are not mere illustrations of a further point. They contribute to the philosophical argument that the author proposes. So if fictions can be thought experiments, then it is possible that fiction can do philosophy both in virtue of and in *spite* of the limitations of its medium (Wartenburg 2006 & 2007).

Tamar Gendler states that "To perform a thought experiment is to reason about an imaginary scenario with the aim of confirming or disconfirming some hypothesis or theory" (2002, 388). Different philosophers use different kinds of thought experiments in order to bolster their own view or undermine their opponent's. For example, a thought experiment can be used as a counterexample to a philosophical thesis, principle, or idea (Wartenberg 2007). Examples include the Gettier cases, Judith Jarvis Thomson's trolley and footbridge dilemmas, and Frank Jackson's Mary (Jackson 1982). A thought experiment may also be used to establish the possibility of a philosophical claim. Descartes' first Meditation establishes the possibility that our senses can deceive us and that we are the victim of a dreaming or evil demon hypothesis (*ibid*, 60). Thought experiments can also confirm a theory. Wartenberg cites Plato's long discussion of a just state as a thought experiment that explains and justifies his theory of a just soul. As we saw in chapter 3, Arthur Danto's Testadura example in "The Artworld" helps to confirm his notion of the 'is' of

aesthetic identification (Danto 1964). Wartenberg also argues that thought experiments can demonstrate the impossibility of a claim, as in Quine's "gavagai!" case, which is intended to show the indeterminacy of translation (Quine 1960). Finally, thought experiments can establish necessary connections, such as how Locke's cobbler and the prince example aims to establish the necessary connection between memory and the self.

Our goal, then, is to show how films (or any fiction) can be or contain thought experiments in one of these ways. Consider *The Walking Dead* "walker-in-the barn" thought experiment from the beginning of chapter 6, based on the episode "Pretty much dead already." We might interpret that as a thought experiment concerning the nature of our moral judgments. Or, my *Boondock Saint*'s example from chapter 1 could be interpreted as a thought experiment concerning the intuitive likelihood of the DAV. Both qualify as thought experiments according to Gendler's definition: they ask viewers to consider a fictional example that attempts to confirm or disconfirm some theory or philosophical idea.

These examples are not unique. Several of Christopher Nolan's films make "metafilm" critiques that could be interpreted as thought experiments; *Inception* illustrates external world skepticism, but also challenges that notion that films must have a tidy conclusion that answers every question it raises (see Jensen 2010). Carroll (2009) argues that the way in which Leonard's experiences are presented in *Memento* comment on film narration in general. As Russell pointed out, Woody Allen's *Crimes and Misdemeanors* can be interpreted as an immoralist challenge to Plato's virtue theory, just like Glaucon's Myth of Gyges in *The Republic*. Dostoevsky's *The Brothers Karamazov* lives up to this philosophical challenge as well, presenting an argument in favor of Ivan's concept of "everything is permitted," then arguing against *that* position with the story of Father Zossima (as well as Ivan's fate at the end of the novel).

However, not every text or film that incorporates philosophical themes counts as a thought experiment. Terence Malick's films *The Thin Red Line*, *The New World*, and *Tree of Life* each present a distinction between Nature and Grace that may engender philosophical thought. I would hesitate to count them as doing philosophy, though, partly because the films are quite vague in terms of presenting an argument (contrast with the above examples which explicitly mention philosophical themes). It's possible that these films are designed to inspire the viewer as opposed to present a philosophical concept or argument.[19]

If at least some fictions can do philosophy, then I argue that it is possible to *learn* from them. The "walker in the barn" case may teach us something about how we should treat other people, if we are responsive to it. *Inception* may teach us something about skepticism and how we interpret film narratives. As we've seen, *Crimes and Misdemeanors* may challenge us to reconsider Plato's notion of justice in *The Republic*.

We lack direct empirical evidence to support the idea that we can learn from fictional thought experiments. However, it seems obvious that we can gain knowledge from philosophy, including philosophical thought experiments. After all, we can learn something conceptual from traditional philosophical thought experiments: Putnam's Twin Earth, Plato's Cave, Jackson's Mary, and Thomson's trolley and footbridge problems are cases in point. These thought experiments are intended to teach us something about metaphysics, knowledge, the mind, and ethics by either presenting positive claims or rejecting others. There is no direct empirical evidence for the claim that reading these works makes us more psychologically or morally sensitive. Consider Shaun Nichols' point that even if psychopaths were to read great works of moral

[19] Malick studied under Stanley Cavell as an undergraduate philosophy major at Harvard. Cavell was one of the original supporters of the idea that films can be philosophy, or at least *philosophical*. See Cavell 1971.

philosophy, it is very unlikely that they would become more morally sensitive (Nichols 2004a).We take it for granted that studying these philosophical works will teach the thoughtful reader about the subject at hand. If we can learn from traditional thought experiments, then it does not seem like there is good reason, in principle, why we cannot learn from fictional thought experiments.

5. Immoral learning

I want to return to Plato's point that fictions can engender immoral practices and values in unsuspecting audiences. This is also Hume's concern in "Of the Standard of Taste": if immoral actions and characters that are not condemned in a work then an unsuspecting audience may take an immoral attitude to be morally acceptable in his real-life (Hume 1757/1994).

Both Plato and Hume worry that consumers of fiction can be easily manipulated into believing that what they witness in a fiction bears, in some way, on reality. But does this actually happen? Recall my discussion of the puzzle of imaginative resistance. Sometimes audiences fail to accept a proposition as true in the fictional world simply because they disagree with the same proposition in the actual world. I have argued that this sometimes occurs when audiences worry that going along with fictional immorality will make themselves more likely to be immoral.

I also suggested that resistance to portrayals of fictional immorality is the result of negative emotional reactions to immoral acts. As we saw with the SDP, it is possible that we do not harshly judge a character like Dexter Morgan because we are drawn to him in some way. But that does *not* necessarily mean that we accept it to be true, even imaginatively, that his immoral actions are

acceptable in either our world or in the world of the fiction. Moreover, I think that portrayals of immorality in artworks generally serve some cognitive or aesthetic purpose. As such, they can present cases of learning from immoral actions. Ideally, acknowledging the cognitive benefits of portrayals of immorality should quell the worries against some fictions.

Consider the *Mad Men* "Mystery Date" example from the previous chapter. This episode portrayed acts of sexual violence towards women and a general tendency to objectify women for men's desires. I argued, though, that this episode probably used the immoral treatment of women to make some point: to cast a light on how women are used in the work place, how casual sexual violence might sometimes seem, and the power struggles that women still face on a daily basis. Fictions often portray immoral characters or actions that are not condemned in the narrative, or at least not explicitly so. I argued that these portrayals of immorality are quite often supposed to make some cognitive or aesthetic point. That is, the audience is supposed to take the portrayal of immorality as an opportunity for moral reflection and learning.

Here's another example. You have now just finished watching Quentin Tarantino's latest film, *Django Unchained*. Some critics have blasted this film for its overt racist elements (Spike Lee is one outspoken critic of the film). There is one scene in particular that seems stands out in terms of its negative portrayal of African Americans. Django, a slave, has agreed to help the white Dr. Schulz with his bounty-hunting after the doctor promises to free him from slavery and help him to find his wife, Bromhilda. Dr. Schulz gives Django some money and tells him that he can buy whatever new clothes he would like. The next scene shows Django, on horseback, proudly sporting his new frock—an absurdly fancy suit made of bright-blue velvet and an abundance of frilly lace. You laugh.

Why? What's so funny about this scene? Well, the clothes, obviously. But is the source of

amusement simply the ridiculousness of the suit itself? I think it is more than that. It seems like this scene is funny because it is a *former slave*—a black man—who wears the stereotypical get-up of a wealthy white man. There is an incongruity between what we expect from Django and what we actually see (Morreall 1986).

One could straightforwardly interpret this scene as Tarantino poking fun at a former slave who ignorantly assumes that he can pull off the garb of a powerful white man. But there is another, probably better, interpretation of this scene. Along with the amusement that one might experience, a viewer might also feel strangely uncomfortable by his or her own amusement at Django's new clothes. I think that this is the point of many of the racist elements in *Django Unchained*: we are *supposed* to laugh *and then* feel uncomfortable about our own amusement. Tarantino seems to be pointing out that, despite our firm denial to the contrary, many white audiences still take African Americans as sources of amusement that should fulfill traditional stereotypes, such as Mammy in *Gone with the Wind*. So we find it comical and a bit absurd when they try to break out of those stereotypes.

If we are sensitive to this aspect of the film then, ideally, we will reflect on our own responses and maybe even gain some kind of self-knowledge about our tacit acceptance of harmful racial stereotypes. I take this to be both a cognitive and aesthetic achievement. If *Django* did not present these racist elements in the way that it did, then one might argue that it negatively portrays African Americans and should be condemned. But the immoral portrayals of Django seem to be of both aesthetic and moral value. The worry remains, though, that many people will not get the lesson behind portrayals of immorality. Some of *Django*'s viewers won't make the cognitive point that I have suggested here. They will not realize that they should feel uncomfortable about their laughter—they just laugh. And this might be morally troubling.

I have argued in this chapter that one can learn about morality from fictions. I suggested that our learning from fictions can be either a type of knowhow, propositional knowledge, or experiential knowledge. As the "Mystery Date" and *Django* examples show, it's possible that we can gain these types of knowledge from portrayals of immorality in a work of fiction. Moreover, these portrayals will generally have some kind of meta-cognitive or aesthetic value. But surely, as the last point about Django suggests, there are examples of fiction that do not seem to make a cognitive or aesthetic point, or make a point that is lost on the audience. Should we be worried about them?

Consider the growing concern that violence in the media—including films, TV shows, and video games—leads to real-life violence (see Provios, Kambam & Bender 2013). Intuitively, it seems like exposing oneself to fictional portrayals of extreme violence is, somehow, causally connected to violence in real-life. We might think that there is some explanatory power to the fact that a serial killer's favorite film is *Scarface* and his favorite video game is *Grand Theft Auto*. Others will deny this, citing anecdotal or even empirical evidence that violent films do not directly cause real-life aggression or anti-social personalities. Perhaps the prominent cases of gun violence are the product of other social factors, not watching violent films or TV.

The empirical evidence on the relationship between fictional and real-life violence has so far been inconclusive. In a meta-analysis of studies, the psychologists Haejung Paik and George Comstock (1994) found that there does seem to be a *correlation* between short-term exposure to media violence and actual violence. Likewise, other researchers found that watching excessive amounts of violent TV (more than two hours per weekday) was associated with antisocial behavior in early adulthood (Robertson et al 2013). Despite these results—and more like them—there does not seem to be any direct causal link between media and real-life violence.

Perhaps the worry is that exposing oneself to immoral actions and characters will make one accustomed to bad things. The concern is that violence might become blasé whether it is portrayed positively or not. Perhaps we will not feel the horror towards real-life violence that we should feel if we are accustomed to seeing it in the media. And it is not just *violence* that we should be worried about: consider the potential long-term effects of sexist and racist practices (see Shottenkirk 2013 & Vadas 1987). The portrayal of sexism and racism may reinforce harmful stereotypes that are already pervasive in our society. Maybe, then, portrayals of immorality in artwork should be condemned.

For the sake of argument, let's suppose that cognitive scientists do find a direct causal relation between media violence and real-life aggression or exposure to racist or sexist fictions cause racism and sexism in real life. What then? Should we, in a moment of Platonic fervor, banish all immoral artworks from our society?

I do not have a good response to this challenge. But I think that there are at least two reasons why we should not censor immoral artworks (besides appeals to freedom of choice and freedom of speech). First, it may be that the supposed immoral act in an artwork is trying to point out some immoral feature of our own society, as in my *Django* example. The immoral features of the fiction are intentionally subversive. *The Adventures of Huckleberry Finn* is the perpetual example of this: Huck knows that freeing his friend Jim from slavery is morally wrong in his society. Huck chooses to free Jim anyway because Jim is his friend. Contemporary audiences praise Huck's decision despite the fact that he may have been found blameworthy by some readers during Twain's time. A fiction can highlight the harmful tendencies of our own society *by* portraying them in a fiction. Second, there is also something to be said for being tolerant and open-minded of the moral practices of other people and cultures. For example, suppose that you watch a film that takes place

in an Arabic nation where gender relations are quite different from what they are in your society. Should we condemn the artwork because it portrays moral values that are different from our own? Some viewers will undoubtedly say that we should, but others will disagree.

I do not intend for either of these reasons to be a full justification against the censorship or condemnation of immoral artworks. However, I do think that we should be sensitive to the potential moral lesson that we can draw from portrayals of immorality in some artworks, like *Django*. Plato may not be assuaged, but, nevertheless, I think that there are quite compelling reasons to be tolerant of fictional portrayals of immorality.

Bibliography

Annas, J. (2005). "Comments on John Doris' 'Lack of Character'." *Philosophy and Phenomenological Research*, 71: 636-642.

Aristotle (1947). *Nichomachean Ethics*. In *Introduction to Aristotle*. R.McKeon (trans.), New York: McGraw-Hill, Inc.

Arnold, M.A. (1960). *Emotion and Personality, Vol. 1: Psychological Aspects*. New York: Columbia University Press.

Bacciarelli, M., S. Khemlani, & P.N. Johnson-Laird (2008). "The psychology of moral reasoning." *Judgment and Decision Making*, 3, 3: 121-139.

Baldwin, D. A. & L. J. Moses (1994). "Early Understanding of Referential Intent and Attentional Focus: Evidence from Language and Emotion." In C. Lewis & P. Mitchell (eds.), *Childrens' early understanding of mind: origins and development*. Hillsdale, NJ: Erlbaum Press.

Baljinder, S. & P. Thagard (2003). "Self-Deception and Emotional Coherence." *Minds and Machines* 13: 213-231.

Barnes, A. (1997). *Seeing through Self-Deception*. Cambridge: Cambridge University Press.

Baron-Cohen, S. (1995). *Mindblindedness: An Essay on Autism and Theory of Mind*. Cambridge, Mass: The MIT Press.

Bayne, T. & M. Spener (2010). "Introspective Humility." *Philosophical Issues*, 20: 1-22.

Beaumont, L.R. (2009). "Our emotional brain: defend, then comprehend." *Emotional Competency*. Visited on 18, November 2013. <http://www.emotionalcompetency.com/ebrain.htm>

Bell, C. (1914/2003). "The Aesthetic Hypothesis." In C. Harrison & P. Wood (eds.) *Art in Theory: 1900-2000*. Malden, MA: Blackwell Publishing, 107-110.

Blackburn, Simon. (1993). *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Blair, R.J.R. (1995). "A Cognitive Development Approach to Morality: Investigating the Psychopath." *Cognition*, 57: 1-29.

Ben Ze'ev, A. (2001). *The Subtlety of Emotions*. Cambridge, MA: The MIT Press.

Bermudez, J.L. (2005). *Philosophy of Psychology*. New York: Routledge.

Berridge, K. C. & P. Winkielman.(2003). 'What is an unconscious emotion? (The case for unconscious liking)'. *Cognition & Emotion* 17: 181-211.

Block, N. (1978). "Troubles with Functionalism." *Minnesota Studies in Philosophy of Science*, 9: 251-325.

--(1995). "On a confusion about a function of consciousness." *Behavioral and Brain Sciences*, 18: 227-287.

--(2005). "Alva Noë: Action in perception." *Journal of Philosophy*, 102.

Botvinick, M. & Cohen, J. (1998). "Rubber hands 'feel' touch that eyes see." *Nature* 391 (6669): 756.

Bower, G.H. & D.G. Morrow (1990). "Mental models in narrative comprehension." *Science* 247.4938: 44+.

Brogaard, B. (2013). Do we perceive natural kind properties? *Philosophical Studies* 162: 35-42.

Byrne, A. (2001). Intentionalism defended. *The Philosophical Review*, 110: 199-240.

Camp, E. (unpublished manuscript). "Perspectives in Imaginative Engagement with Fiction."

--(2009). “Two Varieties of Literary Imagination: Metaphor, Fiction, and Thought Experiments,” *Midwest Studies in Philosophy: Poetry and Philosophy:* 107-130.
Carruthers, P. (1996a). “Simulation and self-knowledge: a defence of theory theory.” In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 22-38.
--(1996b). “Autism as mind-blindness: an elaboration and partial defence.” In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 257-273.
--(2006). *The architecture of mind*. Oxford: Oxford University Press.
--(2010). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.
--(2013). “Mindreading in infancy.” *Mind and Language*, 28:141-172.
Cavell, S. (1971). *The World Viewed*. New York: Viking Press.
Carroll, N. (1990). *The Philosophy of Horror, or Paradoxes of the Heart*. New York: Routledge.
--(1997). “Fiction, Non-Fiction, and the Film of Presumptive Assertion: A Conceptual Analysis.” In *Film Theory and Philosophy*, (eds) R. Allen & M. Smith, New York: Oxford University Press, 173-202.
--(2001a). “Art, Narrative, and Moral Understanding.” In N. Carroll (ed) *Beyond Aesthetics: Philosophical Essays*. Cambridge: Cambridge University Press, 270-293.
--(2001b). “Moderate Moralism.” In N. Carroll (ed) *Beyond Aesthetics: Philosophical Essays*. Cambridge: Cambridge University Press, 293-316.
--(2003). “Art and Mood: Preliminary Notes and Conjectures.” *The Monist*, 86, 521-555.
--(2004). “Sympathy for the Devil.” In R. Greene & P. Vernezze (eds.). *The Sopranos and Philosophy*, Chicago: Open Court, 121-136.
--(2006). “Philosophizing through the Moving-Image: The Case of “Serene Velocity.” *The Journal of Aesthetics and Art Criticism*, Special Issue: Thinking through Cinema: Film as Philosophy, 64, 173-185.
--(2008). *The Philosophy of Motion Pictures*, Malden, MA: Blackwell Publishing.
--(2009). “*Memento* and the phenomenology of comprehending motion picture narration.” In A. Kania (ed.) *Memento*. New York: Routledge, 127-166.
--(2010). “Movies, the Moral Emotions, and Sympathy.” *Midwest Studies in Philosophy, Special Issue: Film and the Emotions*, XXXIV: 1-19.
--(2011). “On some affective relations between audiences and the characters in popular fictions.” In *Empathy: Philosophical and Psychological Perspectives*, A. Coplan & P. Goldie (eds.), Oxford: Oxford University Press, 162-184.
Clark, A. (2000). *A Theory of Sentience*, Oxford: Oxford University Press.
Clay, Z. & M. Iacoboni (2011). “Mirroring fictional others.” In *The Aesthetic Mind: Philosophy and Psychology*, E. Schellekens & P. Goldie (eds.). Oxford: Oxford University Press, 313-331.
Cohen, L. J. (1989). “Belief and Acceptance.” *Mind*, 98: 367-89.
Coleridge, S., T. (1817/1985). *Biographia Literaria.* In *Samuel Taylor Coleridge*, H.J. Jackson (ed.), Oxford: Oxford University Press.
Currie, G. (1988). Photography, painting, and perception. *Journal of Aesthetics and Art Criticism*, 49: 23-29.
--(1990). *The Nature of Fiction*, Cambridge: Cambridge University Press.

--(1995).“The Moral Psychology of Fiction.” *Australasian Journal of Philosophy*, 73: 250-259.
--(1996). “Simulation –theory, theory-theory and the evidence from autism” In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 242-256.
--(1997). “The Paradox of Caring: Fiction and the Philosophy of Mind.” In M. Hjort & S. Laver (eds) *Emotion and the Arts,* Cary, NC: Oxford University Press, 63-77.
--(1999a). “Narrative Desire.” In *Passionate Views: Film, Cognition, and Emotion*, (eds.) Carl Plantinga and Greg. Smith. Baltimore: The John Hopkins University Press, 183-199.
--(1999b). “Visible Traces: Documentary and the Contents of Photographs.” *The Journal of Aesthetics and Art Criticism* 57: 285-97.
--(2002). “Desire in Imagination.” In *Conceivability and Possibility*. Tamar Szabó Gendler and John Hawthorne (eds.). Oxford: Clarendon Press, 201-222.
--(2013). “Does Great Literature Make Us Better?” In *The New York Times*, Opinionator.
--(in progress). “Distorting Mirrors.”
Currie, G. & I. Ravenscroft. (2002). *Recreative Minds*. Oxford: Oxford University Press.
Cushman, F., L. Young, & J. Greene (2010). “Multi-system Moral Psychology.” In *The Moral Psychology Handbook*, J. M. Doris (ed.), Oxford: Oxford University Press, 47-71.
Cushman, F., L. Young, & M. Hauser (2006). “The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm.” *Psychological Science* 17:1082-1089.
Damasio, A. (1994). *Descartes’ Error*. New York: Harper Collins Publishers.
Damasio, A. R., Tranel, D. & Damasio H. (1990). “Face Agnosia and the Neural Substrates of Memory.” Annual Review of Neuroscience, 13: 89-109.
Danto, A. “The Artworld.” *The Journal of Philosophy*, 61: 571-584.
--(1979). “Moving Pictures.” *Quarterly Review of Film Studies*: 4, 1-21.
--(1981). *The Transfiguration of the Commonplace*. Cambridge, MA: Harvard University Press.
Darley, J. M. & C. D. Batson (1973) "'From Jerusalem to Jericho': A study of situational and dispositional variables in helping behavior." *Journal of Personality and Social Psychology*, 27: 100-108.
Daniels, N. (1979). “Wide Reflective Equilibrium and Theory Acceptance in Ethics.” *The Journal of Philosophy* 76: 256-282.
D’Arms, J. & D. Jacobson (2000a). “Sentiment and Value.” *Ethics*, 110: 722-748.
--(2000b).“The Moralistic Fallacy: On the ‘Appropriateness’ of Emotions.” *Philosophy and Phenomenological Research*, LXI: 65-90.
--(2005). “Sensibility Theory and Projectivism.” In *The Oxford Handbook of Ethical Theory*. David Copp (ed.). Oxford: Oxford University Press.
Davies, D. (2007). *Aesthetics and Literature*. London: Continuum.
Davies, S. (1994). *Musical Meaning and Expression.* Ithaca, NY: Cornell University Press.
Decety, J. & A.N. Meltzoff (2011). “Empathy, imitation, and the social brain.” In *Empathy: Philosophical and Psychological Perspectives*, A. Coplan & P. Goldie (eds.), Oxford: Oxford University Press, 58-81.

Dennett, D. (1997). “True believers: The intentional strategy and why it works.” In J. Haugeland (ed.) *Mind Design II: Philosophy, Psychology, Artificial Intelligence*, Cambridge, MA: MIT Press, 57-79.
Deroy, O. (2013). Object sensitivity versus cognitive penetrability of perception. *Philosophical Studies*, 162: 87-107.
Derrida, J. (1987). *The truth in painting*. Chicago: University of Chicago Press.
De Sousa, R. (2002). “Emotional Truth.” *Proceedings of the Aristotelian Society, Supplementary Volumes*, 76: 247-263.
--(2004). “Emotions: What I Know, What I’d Like to Think I Know, and What I’d Like to Think.” In *Thinking about Feeling*, Robert C. Solomon (ed.), New York: Oxford University Press, 61-75.
Devereaux, M. (2004). “Moral Judgments and Works of Art: The Case of Narrative Literature.” Journal of Aesthetics and Art Criticism, 62: 3-11.
De Vignement, F. (2007). Habeas Corpus: the sense of ownership of one’s body.” *Mind and Language*, 22: 427-449.
--(2011). “A mosquito bite against the enactive approach to bodily experiences.” Journal of Philosophy, 108: 188-204.
Dijksterhuis, A. & Nordgren, L.F. (2006). “A theory of unconscious thought.” *Perspectives on Psychological Science*, 1: 95-109.
Dokic, J. (2012). “Pictures in the Flesh: Presence and Appearance in Pictorial Experience.” *British Journal of Aesthetics*, 52: 391-405.
Doris, J.M. (1998). "Persons, Situations, and Virtue Ethics." *Nous*, 32, 504-530.
--(2002). *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.
Dretske, F. (1995). *Naturalizing the Mind*, Cambridge, MA: MIT Press.
Eckman, P. (1999). “Basic Emotions.” In T. Dalgleish & T. Power (eds.) *The Handbook of Emotion and Cognition*. New York: Wiley, 45-60.
Ellsworth, P. (1994). “William James and Emotion: Is a Century of Fame Worth a Century of Misunderstanding?” *Psychological Review* 101: 222-229.
Evans, G. (1982). *Varieties of Reference*. Oxford: Clarendon Press.
Feagin, S. (1983). “The Pleasures of Tragedy.” *American Philosophical Quarterly*, 20: 95-104.
--(1984). “Some Pleasures of Imagination.” *Journal of Aesthetics and Art Criticism* 43:41-55.
--(2010). “Film Appreciation and Moral Insensitivity.” *Midwest Studies in Philosophy, XXXIV*, 20-33.
--(2011). “Empathizing as simulating.” In *Empathy: Philosophical and Psychological Perspective*s, A. Coplan and P. Goldie (eds.), Oxford: Oxford University Press, 149-161.
Fish, W. (2009). *Perception, hallucination, and illusion*. New York: Oxford University Press.
--(2010). Philosophy of Perception. New York: Routledge.
Fodor, J. (1983). *Modularity of Mind*. Cambridge, MA: The MIT Press.
Foucault, M. (1970). *The Order of Things*. New York: Random House.
Freedburg, D. & V. Gallese. (2007). “Motion, emotion, and empathy in esthetic experience,” *TRENDS in Cognitive Sciences*, 11:197-203.
Gallagher, S. (forthcoming). “The new hybrids: Continuing debates on social perception.”
Gallagher, S. & S. Varga (2014). “Social constraints on the direct perception of emotions and intentions.” *Topoi*, 33: 185-199.

Gaut, B. (1999). "Identification and Emotion in Narrative Film." In C. Plantinga & G.M. Smith (eds) *Passionate Views: Film, Cognition, and Emotion*, Baltimore: The John Hopkins University Press, 200-216.
--(2007). *Art, Emotion and Ethics*. Oxford: Oxford University Press.
Gendler, T. S. (2000). "The Puzzle of Imaginative Resistance," *The Journal of Philosophy*, 97: 55-81.
--(2006). "Imaginative Resistance Revisited." In The Architecture of the Imagination, Shaun Nichols (ed.), Oxford: Oxford University Press, 149-174.
--(2008). "Alief and Belief." *The Journal of Philosophy*, 105: 634-663.
Gendler, T.S. & Kovakovich (2005). "Genuine rational fictional emotions." In *Contemporary Debates in Aesthetics and the Philosophy of Art*. Malden, MA: Blackwell, 242-253.
Gettier, E. (1963). "Is Justified True Belief Knowledge?" *Analysis* 23, 121-123.
Gibson, J. (2003). "Between Truth and Triviality." *British Journal of Aesthetics*, 43: 224-237.
Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.
Gilovich, T. & Griffin, D. (2002). "Heuristics and biases: Then and now." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich & D. Griffin (eds.), Cambridge: Cambridge University Press, 1-17.
Gigerenzer, G. (2008). "Moral Intuition=Fast and Frugal Heuristics?" In *Moral Psychology, Vol 2, The Cognitive Science of Morality: Intuition and Diversity*, W. Sinnott-Armstrong (ed.), Cambridge, MA: The MIT Press, 1-26.
Gigerenzer, G., Czerlinski, J. & Martignon, L. (2002). "How good are fast and frugal heuristics?" In *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich & D. Griffin (eds.), Cambridge: Cambridge University Press, 559-581.
Goldman, A. (1989). "Interpretation psychologized." *Mind & Language*, 4: 161-185.
--(1993). "The psychology of folk psychology." *Behavioral and Brain Sciences*, 16: 15-28.
--(2006). "Imagination and Simulation in Audience Responses to Fiction." In *The Architecture of the Imagination: new essays on pretence, possibility, and fiction.* Shaun Nichols (ed.). Oxford: Clarendon Press.
--(2008). *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. Oxford: Oxford University Press.
Goldie, P. (2000). *The Emotions*. New York: Oxford University Press.
--(2004)."Emotion, Feeling, and Knowledge of the World." In *Thinking about Feeling*. Robert C. Solomon (ed.). Oxford: Oxford University Press, 91-106.
--(2012). *The Mess Inside*. Oxford: Oxford University Press.
Gombrich. E. (1960). *Art and Illusion*. The A.W. Mellon Lecture Series in the Fine Arts. Bollingen Series XXXV: 5. Princeton, NJ: Princeton University Press.
Goodman, N. (1976). *Languages of Art*. Indianapolis, IN: Hackett Publishing Company, Inc.
Gopnik, A. (1993). "How we read our own minds: The illusion of first-person knowledge of intentionality." *Behavioral and Brain Sciences*, 16, 1-14.
Gopnik, A. & Schulz, L. (2004). "Mechanisms of theory-formation in young children." *Trends in Cognitive Sciences*, 8: 371-377.
Gordon, R. (1986). "Folk Psychology as Simulation." *Mind and Language* 1: 158-171.
--(1987). *The Structure of Emotions*. Cambridge: Cambridge University Press.
--(1996). " 'Radical' simulationism" In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 11-21.

Graesser, A.C., Singer, M. & T. Trabasso (1994). "Constructing inferences during narrative text comprehension." *Psychological Review*, 101: 371-395.
Greenberg, C. (1940/2003). "Towards a New Laocoon." In C. Harrison & P. Wood (eds.) *Art in Theory: 1900-2000*. Malden, MA: Blackwell Publishing, 562-568.
Greene, J. (2007). "The Secret Joke of Kant's Soul." In *Moral Psychology, Vol 3, The Neuroscience of Morality*, W. Sinnott-Armstrong (ed), Cambridge, MA: The MIT Press, 35-80.
Greene, J., R. Sommerville, L. Nystrom, J.M. Darley, & J. Cohen (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293: 2105-2108.
Greene, J. & J. Haidt (2002). "How (and where) does moral judgment work?" *TRENDS in Cognitive Sciences*, 6: 517-523.
Greenwood, J. (1991). *Relations, representations*. London: Routledge.
Griffiths, P. (1997). *What Emotions Really Are*. Chicago: Chicago University Press.
--(2002). "Appraisal and Machiavellian Emotion." The Proceedings of the Emotion, Evolution, and Rationality Conference. Accessed 18 November 2013. < http://philsci-archive.pitt.edu/667/>
--(2004). "Is Emotion a Natural Kind?" In *Thinking about Feeling*. Robert C. Solomon (ed.). Oxford: Oxford University Press, 233-249.
Gross, J.L. (2007). *Handbook of Emotion Regulation*. New Graves, T. L., B. Maniscalco & H. Lau. (2010)."Volition and the Function of Consciousness," *Conscious Will and Responsibility*, Ed. Walter Sinnott-Armstrong and Lynn Nadel, New York: Oxford University Press.
Haggard, P. (2005). "Conscious intentions and motor cognition." *TRENDS in Cognitive Science*, 9: 290-295.
Haggard, P. & B. Libet. (2001). "Conscious Intention and Brain Activity," *Journal of Consciousness Studies*, 8, 47-63.
Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, 108: 814-834.
Haidt, J. & F. Bjorkland (2008). "Social Intuitionists Answer Six Questions about Moral Psychology." In *Moral Psychology, Vol 2, The Cognitive Science of Morality: Intuition and Diversity*, W. Sinnott-Armstrong (ed.), Cambridge, MA: The MIT Press, 181-218.
Harman, G. (1990). The intrinsic quality of experience. In J.E. Tomberlin (ed.), *Philosophical Perspectives 4: Action Theory and the Philosophy of Mind*. Atascadero, CA: Ridgeview, 31-52.
--(1996). "Moral Relativism," in G. Harman and J.J. Thompson (eds.) *Moral Relativism and Moral Objectivity*, Cambridge MA: Blackwell Publishers, 3–64.
--(2000). "The Nonexistence of Character Traits." *Proceedings of the Aristotelian Society*, 100: 223-226.
Harman, G., Mason, K. & W. Sinnott-Armstrong (2010). "Moral Reasoning." In J. Doris (ed.) *The Moral Psychology Handbook*, New York: Oxford University Press.
Harris, P (2000). *The Work of the Imagination*. Oxford: Blackwell Publishers, Ltd.
Hauser, M. (2006). *Moral minds: How nature designed a universal sense of right and wrong*. New York: Ecco Press/Harper Collins.
Haybron, D.M. (2007). "Do we know how happy we are? On some limits of affective introspection and recall." *Nous*, 41:3, 394-428.

Heal, J. (1996). "Simulation, theory, and content." In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 75-89.
Hibberd, J. (2013). "'Game of Thrones' author George R.R. Martin: Why he wrote The Red Wedding." Accessed 6 June, 2013. <http://insidetv.ew.com/2013/06/02/game-of-thrones-author-george-r-r-martin-why-he-wrote-the-red-wedding/>.
Hochberg, J. & V. Brooks (1962). Pictorial recognition as an unlearned ability? *American Journal of Psychology*, 75, 624-8.
Hoffman, M. (2010). "Empathy and pro-social behavior" In M. Lewis & J.M Haviland-Jones (eds.) *Handbook of Emotions* (3rd ed.), New York: The Guildford Press, 440-455.
Hopkins, R. (2008). What do we see in film? *The Journal of Aesthetics and Art Criticism*, 66: 149-159.
Huebner, B., S. Dwyer, & M. Hauser (2008). "The role of emotion in moral psychology." *TRENDS in Cognitive Sciences* 13: 1-6.
Hume, D. (1757/ 1994). "Of the Standard of Taste." In S. D. Ross (ed.) *Art and its Significance*, Albany, NY: State University of New York Press.
Hurka, T. (2001). *Virtue, Vice, and Value*. Oxford: Oxford University Press.
Iser, W. (1974). *The implied reader: Patterns in communication in prose fiction from Bunyan to Beckett*. Reprinted in *Reader-response criticism: From formalism to post-structuralism*, J.P. Tompkins (ed.), Baltimore: The John Hopkins Press, 50-69.
Jackson, F. (1982). "Epiphenomenal Qualia." *Philosophical Quarterly*, 32, 27-36.
James, W. (1890/2007). *Principles of Psychology*. New York: Cosimo.
Jenson, J. (2010). "Christopher Nolan on his 'last' Batman movie, an 'Inception' videogame, and that spinning top." *Entertainment Weekly*, Nov. 30th.
Johnson-Laird, P. (2008). "Mental Models and Deductive Reasoning." In Reasoning*: Studies in Human Inference and its Foundations*, L. Rips & J. Adler (eds.), Cambridge: Cambridge University Press, 206-222.
Kant, I. (1785/1959). *Foundations of the Metaphysics of Morals*. L. W. Beck (trans.), New York: Library of Liberal Arts.
(1797/1996). *The Metaphysics of Morals*. M.J. Gregor (trans.). Cambridge: Cambridge University Press.
Kidd, D.C. & E. Castano (2013). "Reading literary fiction improves theory of mind." *Science*, 342: 377-380.
Kieran, M. (2006). "Art, Morality, and Ethics: On the (Im)Moral Character of Art Works and Inter-Relations to Artistic Value." *Philosophy Compass* 1/2: 129-143.
--(2010). "Emotions, Art, and Immorality." In P. Goldie (ed.) *The Oxford Handbook of Philosophy of Emotion*, Oxford: Oxford University Press, 681 - 704.
Kivy, P. (1999). "Feeling the Musical Emotions." *British Journal of Aesthetics*, 39: 1-13.
Kosslyn, S. (1997). "Neural Systems Shared by Visual Imagery and Visual Perception: A Positron Emission Tomograpy Study." *Neuro-Image*, 6: 320-34.
Kosslyn, S., Alpert, N., Thompson, W., Maljokovic, V. Weise, S., Chabris, C., Hamilton, S., Rauch, S., & Buonnano, E. (1993). "Visual Mental Imagery Activates Topographically Organized Visual Cortex: PET Investigations." *Journal of Cognitive Neuroscience*, 5: 263-87.
Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.
--(2011). "Vacuous Names and Fictional Entities." In *Philosophical Troubles*, Vol 1. New York: Oxford University Press.

--(2013). *Reference and Existence: The John Locke Lectures*. Oxford: Oxford University Press.
Kohlberg, L. (1981). *Essays on Moral Development, Vol 1*. New York: Harper & Row.
Korsgaard, C.M. (1986). "Skepticism about Practical Reason." *The Journal of Philosophy* 83: 5-25.
Lamarque, P. (1981). "How Can We Fear and Pity Fictions?" *British Journal of Aesthetics*, 21: 91-304.
--(2009). *The Philosophy of Literature*. Malden, MA: Blackwell Publishing.
--(2011). "On keeping psychology out of literary criticism." In *The Aesthetic Mind: Philosophy and Psychology*, E. Schellekens & P. Goldie (eds.), Oxford: Oxford University Press, 299-312.
Lamarque, P. & S.H. Olsen (1997). *Truth, Fiction, Literature*. Oxford: Clarendon Press.
Lavelle, J.S. (2012). "Theory-theory and the direct perception of mental states." *Review of Philosophy and Psychology*, 3: 213-230.
Lazarus, R. (1991). *Emotion and Adaptation*. New York: Oxford University Press.
LeDoux, J. E. (1996). *The Emotional Brain*. New York: Simon & Schuster.
--(2012)"Rethinking the Emotional Brain." *Neuron, 73*, 653-676.
LeDoux, J.E. & E. A. Phelps (2008). "Emotional Networks in the Brain." In *Handbook of Emotions*, 3rd edn. Michael Lewis, Jeanette M. Haviland-Jones & Lisa Feldman Barrett (eds.). New York: The Guilford Press, 159-179.
Leslie, A. M. (2000). "'Theory of mind' as a mechanism of selective attention." In M. Gazzangiga (ed.) *The new cognitive neurosciences*, Cambridge, MA: MIT Press, 2nd ed., 1235-1247.
Lewis, D. (1972). "Psychophysical and theoretical identifications." *Australasian Journal of Philosophy*, 50: 249-258.
--(1978). "Truth in fiction." *American Philosophical Quarterly*, 15: 37-46.
Libet, B. (1985). "Unconscious cerebral initiative and the role of conscious will in voluntary action." *Behavioral and Brain Sciences*, 8: 529-566.
Livingston, P. (1997). "Cinematic Authorship." In R. Allen & M. Smith (eds.) *Film theory and philosophy*, New York: Oxford University Press, 132-148.
--(2006). "Theses on Cinema as Philosophy." *The Journal of Aesthetics and Art Criticism*, Special Issue: Thinking through Cinema: Film as Philosophy, 64: 11-18.
Lopes, D. (2003). "The Aesthetics of Photographic Transparency." *Mind,* 112: 335-348.
--(2005). *Sight and Sensibility.* Oxford: Oxford University Press.
Lyons, J. (2005). "Perceptual Belief and Nonexperiential Looks." *Philosophical Perspectives*, 19: 237-256.
Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong*. Penguin Books.
Mandelbaum, E. (2014). Thinking is Believing. *Inquiry* 57: 55-96.
Marraffa, M. (2013). "Theory of mind." *Internet Encyclopedia of Philosophy*, <http://www.iep.utm.edu/theomind/>
Martin, M.G.F. (2004). "The Limits of Self-Awareness," *Philosophical Studies*, 120: 37–89.
Matravers, D. (1991). "Who's afraid of Virginia Woolf?" *Ratio*, 1, 25-37.
--(2014). *Fiction and Narrative*. Oxford: Oxford University Press.
McDowell, J. (1978). "Are Moral Requirements Hypothetical Imperatives?" *Proceedings of the Aristotelian Society*, supp. Vol. 52: 13-29.

--(1984). "Values and Secondary Qualities," in *Morality and Objectivity*, T. Honderich and Paul Kegan (eds.), London: Routledge, 110–129.
--(1994). *Mind and World*, Cambridge, MA: Harvard University Press.
--(1998). *Mind, Value, and Reality*. Cambridge, MA: Harvard University Press.
Meinong, A. (1904/1981). "Theory of objects." In R. Chislholm & (ed.) *Realism and the Background of Phenomenology*, I. Levi, D. B. Terrell, and Roderick Chisholm (trans.). Atascadero, CA: Ridgeview, 76-117.
Mele, A. (2001). *Self-Deception Unmasked*. Princeton, NJ: Princeton University Press.
Mello, L. & Villares, J. (1997). "Neuroanatomy of the Basal Ganglia." *Psychiatric Clinics of North America*, 20: 691-704.
Messaris, P. (1994). *Visual literacy: Image, mind and reality*. Boulder, CO: Westview Press.
Mill, J.S. (1861/2002). *Utilitarianism*. 2nd Ed. Indianapolis, IN: Hackett Publishing Company, Inc.
Miller, A. (1996). "An objection to Smith's argument for internalism." *Analysis* 56: 169-174.
Milton, J. (1667/2007). *Paradise Lost*. Ann Arbor, MI: Borders Classics.
Moors, A., Ellsworth, P.C. Scherer, K.R., & Frijda, N. (2013). "Appraisal Theories of Emotion: State of the Art and Future Developmoent." *Emotion Review*, 5: 119-124.
Moran, R. (1994). "The Expression of Feeling in Imagination." *The Philosophical Review*, 103: 75-106.
Nichols, S. & S. Stich (2003). *Mindreading*. Oxford: Oxford University Press.
--(2001). *Authority and estrangement: an essay in self-knowledge*. Princeton, NJ: Princeton University Press.
Morreall, J. (1986). *The Philosophy of Laughter and Humor*. New York: State University of New York Press.
Morton, A. "Empathy for the Devil." In *Empathy: Philosophical and Psychological Perspectives*, A. Coplan & P. Goldie (eds.), Oxford: Oxford University Press, 318-330.
Murdoch, I. (1970). *The Sovereignty of Good*. London: Routledge.
Nabokov, V. (1970). *The Annotated Lolita*. Ed. A. Appel, Jr. New York: McGraw-Hill Company.
Neill, A. (1991). "Fear, fiction, and make-believe" *The Journal of Aesthetics and Art Criticism*, 49: 47-56.
--(1993). "Fiction and the emotions." *American Philosophical Quarterly*, 1: 1-13.
Nichols, S. (2002). "How Psychopaths Threaten Moral Rationalism, or Is it Irrational to be Amoral?" *The Monist* 85: 285-304.
--(2004a). *Sentimental Rules*. New York: Oxford University Press.
--(2004b). "Imagining and believing: The promise of a single code." *The Journal of Aesthetics and Art Criticism*, Special issue on Art, Mind, and Cognitive Science, 62, 129-139.
--(2008). "Sentimentalism Naturalized." *Moral Psychology, Vol 2, The Cognitive Science of Morality: Intuition and Diversity*, W. Sinnott-Armstrong (ed.), Cambridge, MA: The MIT Press, 255-274.
Nichols, N., Stich, S., Leslie, A., & D. Klein (1996). "Varieties of off-line simulation." In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 39-73.
Nisbett, R. & T. Wilson (1977). "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review*, 84, 231–259.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.
--(2006). "Experience without the Head." In Tamar Gendler and John Hawthorne (eds) *Perceptual Experience*, Oxford: Clarendon Press, 411-433.
Noë, A. & O'Regan, J.K. (2001). "A sensorimotor account of visual consciousness." *Behavioral and Brain Sciences*, 24: 939-973.
Nucci, L.P. (1982). "Conceptual Development in the Moral and Conventional Domains: Implications for Values Education." *Review of Education Research* 52: 93-122.
Nussbaum, M. (1990). *Love's Knowledge*. New York: Oxford University Press.
--(2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge, Cambridge University Press.
O'Cravan, K.M. & Kanwisher, N. (2000). "Mental Imagery of Faces and Places Activates Corresponding Stimulus-specific Brain Regions." *Journal of Cognitive Neuroscience*, 12: 1013-23.
O'Shaunessey, B. (2000). *Consciousness and the World*, Oxford: Clarendon.
Paik, H. & G. Comstock (1994). "The Effects of Television Violence on Antisocial Behavior: A Meta-Analysis." *Communication Research*, 21: 516–46.
Parsons, L., Gabrieli, J., Phelps, E., & Gazzaniga, M. (1998). "Cerebrally Lateralized Mental Representations of Hand Shape and Movement." *Journal of Neuroscience*, 18: 6539-48.
Paxton, J., L. Ungar, & J. Greene (2011). "Reflection and Reasoning in Moral Judgment." *Cognitive Science*: 1-15.
Peacocke, C. (1992). *A Study of Concepts*, Cambridge, MA: MIT Press.
Perner, J. (1996). "Simulation as explicitation of prediction-implicit knowledge about the mind: arguments for a simulation-theory mix" In *Theories of theories of mind*, P. Carruthers & P.K. Smith (eds.), Cambridge: Cambridge University Press, 90-104.
Plantinga, A. (1974). *The Nature of Necessity*. Oxford: Oxford University Press.
Plantinga, C. (1999). "The Scene of Empathy and the Human Face on Film." In *Passionate Views: Film, Cognition, and Emotion*, (eds.) Carl Plantinga and Greg. Smith. Baltimore: The John Hopkins University Press, 239-255.
Plato (1985). *The Republic*. Trans. R.W. Sterling & W. C. Scott. New York: W.W. Norton & Company.
Priest, G. (1997). "Sylvan's Box: A Short Story and Ten Morals." *Notre Dame Journal of Formal Logic*, 38: 573-582.
Prinz, J.J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, Mass: The MIT Press.
--(2004a). *Gut Reactions*, Oxford: Oxford University Press.
--(2004b). "Embodied Emotions." In Robert C. Solomon (ed.) *Thinking about Feeling*, New York: Oxford University Press.
--(2006). "The Emotional Basis of Moral Judgments." *Philosophical Explorations* 9: 29-43.
--(2007a). *The Emotional Construction of Morals*. Oxford: Oxford University Press.
--(2007b). "Is morality innate?" In *Moral Psychology, Vol 1, The Evolution of Morality: Adaptations and Innateness*, W. Sinnott-Armstrong (ed.), Cambridge, MA: The MIT Press, 367-406.
--(2011). "Emotion and Aesthetic Value." In E. Schellekens & P. Goldie *The Aesthetic Mind.* Oxford: Oxford University Press.

Prinz, J.J. & S. Nichols (2010). "Moral Emotions." In *The Moral Psychology Handbook*, J. Doris (ed.) New York: Oxford University Press.
Provios, V.K., P.R., Kamben, & H.E. Bender (2013). "Does media violence lead to the real thing?" *The New York Times*, 23, Aug. 2013. <http://www.nytimes.com/2013/08/25/opinion/sunday/does-media-violence-lead-to-the-real-thing.html?_r=0>
Putman, H. (1978). "Literature, Science, and Reflection." In *Meaning and the Moral Sciences*, London: Routledge.
--(1981). *Reason, Truth, and History*. Cambridge: Cambridge University Press.
Pylyshyn, Z. (1999). Is vision continuous with cognition? The case or cognitive impenetrability of visual perception. *Behavioral and Brain Sciences,* 22: 341-365.
Quilty-Dunn, J. (*forthcoming*). Believing our eyes: the role of false belief in the experience of cinema. *British Journal of Aesthetics*.
Quine, W.V.O. (1951). "Two Dogmas of Empiricism." *Philosophical Review* 60: 20-43.
--(1953). "On what there is." *From a Logical Point of View*. Cambridge, MA: Harvard University Press.
--(1960). *Word and Object*. Cambridge, MA: The MIT Press.
Rachels, J. (2003). *The Elements of Moral Philosophy*, 4th edn. New York: McGraw-Hill Company.
Radford, C. (1975). "How can we be moved by the fate of Anna Karenina?" *Proceedings of the Aristotelian Society,* 49: 67-93.
Renero, A. & Tullmann, K. (in preparation). "Experiencing sounds in time."
Richardson, C. (2009). "Phantom Limb Pain; Prevalence, Mechanisms and Associated Factors." In *Amputation, prosthesis use, and phantom limb pain: an interdisciplinary perspective*, C. Murray (ed), 137-156.
Robertson, L., H.M. McAnally, & R.J. Handcox (2013). "Children and adolescent television viewing and antisocial behavior in early adulthood." *Pediatrics*, 131: 431-446.
Robinson, J. (2005). "Emotion: Biological Fact or Social Construction?" In *Thinking about Feeling*, Robert C. Solomon (ed.), New York: Oxford University Press, 28-43.
--(2005). *Deeper than Reason*. New York: Oxford University Press.
--(2010). "Emotional response to music: What are they? How do they work? And are they relevant to aesthetic appreciation?" In P. Goldie (ed.) *The Oxford Handbook of Philosophy of Emotion*, Oxford: Oxford University Press, 651-680.
Roedder, E. & G. Harman (2010). "Linguistics and Moral Theory." In J. Doris (ed.) *The Moral Psychology Handbook*, New York: Oxford University Press.
Rolls, E. (2000). "Orbitofrontal Cortex and Reward." *Cerebral Cortex*, 10: 284-94.
Roseman, I.J. (1996). "Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory." *Cognition & Emotion*, 10, 241 -278.
Roseman, I.J.& Smith, C.A. (2001). Appraisal theory: overview, assumptions, varieties, controversies. New York, NY: Oxford University Press.
Rosenthal, D. (2005). *Consciousness and mind*. New York: Oxford University Press.
--(2008). "Consciousness and its function," *Neuropsychologia*, 46: 829-840.
Ruskin, J. (1870/2006). *Lectures on Art*. Project Gutenburg. Accessed 15 November 2013. http://www.gutenberg.org/files/19164/19164-h/19164-h.htm
Sainsbury, R.M. (2010). *Fiction and Fictionalism*. New York: Routledge.
Russell, Bertrand (1905). "On denoting." *Mind* 14: 479-493.

Russell, Bruce. (2000). "The Philosophical Limits of Film." *Film and Philosophy*, Special Edition: 163-167.
Sabini, J. &M. Silver. (2005). "Lack of Character? Situationism Critiqued." *Ethics*, 115, 535-562.
Salmon, N. (1998). "Nonexistence." *Noûs* 32: 277-319.
Sartre, J.P. (1940/2004). *The Imaginary: A Phenomenological Psychology of the Imagination*. J. Webber (trans.), London and New York: Routledge.
Schachter, S. & Singer, C. (1962). "Cognitve, social, and physiological determinants of emotional states." *Psychological Review*, 69: 379-399.
Schafer-Landau, R. (2003). *Moral Realism: A Defense*. Oxford: Oxford University Press.
--(2008). "Defending Ethical Intuitionism." *Moral Psychology, Vol 2, The Cognitive Science of Morality: Intuition and Diversity*, W. Sinnott-Armstrong (ed.), Cambridge, MA: The MIT Press, 83-96.
Scherer, K.R. (1984). "Emotion as a multi-level process." In P. Shaver (ed.) *Review of Personality and Social Psychology*, 5: 37-63.
--(1999). "Appraisal Theory." In *The Handbook of Emotion and Cognition*, T. Dalgleish & M.J. Power (eds.), Chichester, NY: John Wiley and Sons.
Scherer, K. R., & Shorr, A., & Johnstone, T. (Ed.). (2001). Appraisal processes in emotion: theory, methods, research. Canary, NC: Oxford University Press.
Schnall, S., J. Haidt & G. Clore (2008). "Disgust as embodied moral judgment." *Personality and Social Psychology Bulletin* 34: 1096-1109.
Schroeder, T. & C. Matheson. (2006). "Imagination and Emotion." In *The Architecture of the Imagination: new essays on pretence, possibility, and fiction.* Shaun Nichols (ed.). Oxford: Clarendon Press.
Schroder, T., Roskies, A.L, & S. Nichols (2010). "Moral Motivation." In J. Doris (ed.) *The Moral Psychology Handbook*, New York: Oxford University Press.
Schiffer, S. (1996). "Language-Created Language-Independent Entities." *Philosophical Topics* 24: 149-167.
Schwitzgebel, E. (2008). "The unreliability of naïve introspection." *Philosophical Review*, 117, 245-273.
Scott, A.O. (2013). "A New Year, And a Last Day Alive." *The New York Times*. Accessed on 1 October 2013. <http://www.nytimes.com/2013/07/12/movies/fruitvale-station-is-based-on-the-story-of-oscar-grant-iii.html?_r=0>
Scruton, R. (1981). "Photography and Representation." *Critical Inquiry* 7: 577-603.
Searle, J. (1975). "The Logical Status of Fictional Discourse." *New Literary History* 6: 319-32.
Sellars, W. (1953). "Empiricism and the philosophy of mind." In *Science, perception, and reality*, London and New York: Routledge & Kegan.
Shottenkirk, D. (2013). "GTA5: Kitsch is a funny thing." 2 October 2013 (blog post). < http://denashottenkirk.wordpress.com/2013/10/02/gta5-kitsch-is-a-funny-thing/>.
Siegel, S. (2006). Which properties are represented in perception? In *Perceptual Experience*, T. Szabo Gendler and J. Hawthorne (Eds.), Oxford: Oxford University Press, 481-503.
--(2010). The contents of perception. *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/perception-contents/>.
--(2011). Cognitive penetrability and perceptual justification. *Nous*, 46: 201-222.
Siewert, C. (1998). *The Significance of Consciousness*. Princeton: Princeton University Press.

Simon, D., Pham, L.B., Le, Q.A., & K. Holyoak (2001). "The Emergence of Coherence Over the Course of Decision Making." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27: 1250-1260.

Singer, P. (1995). *How Are We to live?: Ethics in an Age of Self-Interest*. Amherst, NY: Prometheus Books.

Sinnott-Armstrong, W. (2008). "Framing Moral Intuitions." In *Moral Psychology, Vol 2, The Cognitive Science of Morality: Intuition and Diversity*, W. Sinnott-Armstrong (ed.), Cambridge, MA: The MIT Press, 47-76.

Sinnott-Armstrong, W., Young, L. & Cushman, F. (2010). "Moral Intuitions." In *The Moral Psychology Handbook*, J. Doris (ed.), Oxford: Oxford University Press.

Sloman, S.A. (2002). "Two systems of reasoning." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich & D. Griffin (eds.), Cambridge: Cambridge University Press, 379-397.

Slote, M. (2005). "Moral Sentimentalism and Moral Psychology." In *The Oxford Handbook of Ethical Theory*. David Copp (ed.). Oxford: Oxford University Press.

Slovic, P., Finucane, M., Peters, E., & MacGregor D.G. (2002). "The affective heuristic." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich & D. Griffin (eds.), Cambridge: Cambridge University Press, 397-420.

Smith, C.A. & L.D. Kirby (2001). "Toward Delivering on the Promise of Appraisal Theory." In C.A. Smith & I.J. Roseman (eds.) Appraisal theory: overview, assumptions, varieties, controversies. New York, NY: Oxford University Press, 121-138.

Smith, C.A. & R.S. Lazarus (1990). "Emotion and Adaptation." In L.A. Pervin (ed.) *Handbook of Personality*. New York: The Guilford Press, 609-637.

Smith, G.M. (1999). "Gangsters, Cannibals, Aesthetes, or Apparantly Perverse Allegiances." In C. Plantinga & G.M. Smith (eds.) *Passionate Views: Film, Cognition, and Emotion*. Baltimore: The John Hopkins University Press, 217-238.

Smith, M. (1996). "The Argument for Internalism: Reply to Miller." *Analysis* 56:175-184.

--(2008). "The Truth about Internalism." In *Moral Psychology Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* , W. Sinnott-Armstrong (ed), New York: Oxford University Press, 207-215.

Smith, P.D. (1962). "Vladimir Nabokov on his life and work." *The Listener*, 22 November 1962, 856–858.

Smuts, A. (in preparation). "Pleasurably Regarding the Pain of Fictional Others."

Solomon, R. (1973). "Emotions and choice." *The Review of Metaphysics*, 27: 20-41.

--(1993). *The Passions: Emotions and the Meaning of Life.* 2d. Ed. Indianapolis, IN: Hackett.

--(2004). "Emotions, Thoughts, and Feelings: Emotions as Engagements with the World." In *Thinking about Feeling*, Robert C. Solomon (ed). New York: Oxford University Press, 76-89.

Sreenivasan, G. (2002). "Errors about Errors: Virtue Theory and Trait Attribution." *Mind*, 111, 47-68.

--(2013). "The situationist critique of virtue ethics." In D. Russell (ed.) *Cambridge Companion to Virtue Ethics*, Cambridge: Cambridge University Press, 290-314.

Stein, N.L. (1996). "Children's memory for emotional events: Implications for testimony." In K. Pezdek & W. P. Banks (eds.), *The recovered memory/false memory debate*. San Diego, CA: Academic Press, 169-196.

Stein, N.L., T. Trabasso, & M.D. Liwag (2000). "A Goal-appraisal theory of emotional understanding: Implications for development and learning. In In *Handbook of Emotions*, 3rd edn. Michael Lewis, Jeanette M. Haviland-Jones & Lisa Feldman Barrett (eds.). New York: The Guilford Press, 436-457.
Stokes, D. (2014). "Cognitive penetration and the perception of art." *Dialectica*. 1: 1-34.
Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: The MIT Press.
The Actual Freedom Trust, "Our Animal Instinctual Passions in the Human Brain." Visited on 18, November 2013. <http://www.actualfreedom.com.au/library/topics/instincts.htm>
Thomasson, A. (1999). *Fiction and Metaphysics*. Cambridge: Cambridge University Press.
--(2003). "Speaking of fictional characters." *Dialectica* 57: 205-223.
Thomson-Jones, K. (2008). *Aesthetics & Film*. Cambridge: Continuum.
Timpe, K. (2008). "Moral character." *Internet Encyclopedia of Philosophy*. <http://www.iep.utm.edu/moral-ch/>
Toffler, A. (1964). "Playboy interview: Vladimir Nabokov." *Playboy*, January: 35
Tullmann, K. (2010). The Nature of Fictional Film Characters." Master's Thesis, University of Missouri-St. Louis.
Tullmann, K. & Buckwalter (2014). "Does the paradox of fiction exist?" *Erkenntnis* 79: 779-796.
Turiel, E.(1975). "The Development of Social Concepts: Mores, Customs, and Conventions." In *Moral Development: Current Theory and Research*, D.J. DePalma & J.M. Foley (eds.), Hillsdale, NJ: Erlbaum.
Turiel, E., M. Killen, & C. Helwig (1987). "Morality: Its Structure, Functions, and Vagaries." In *The Emergence of Morality in Young Children*, J. Kagan & S. Lamb (eds.), Chicago: The University of Chicago Press, 155-243.
Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.
Tversky, A. & Kahneman, D. (2002). "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich & D. Griffin (eds.), Cambridge: Cambridge University Press, 19-48.
Vadas, M.(1987). "A first look at the pornography/civil rights ordinance: could pornography be the subordination of women?" *The Journal of Philosophy*, 84: 487-511.
Van Leeuwen, N. (2013). "The Meanings of 'Imagine' Part I: Constructive Imagination." *Philosophy Compass*, 8/3: 220-230.
Velleman, D. (2000). *The Possibility of Practical Reason*. New York: Oxford University Press.
Vosniadou, S. (1987). Children and Metaphors. *Child Development,* 58, 870-885.
Walton, K. (1973). "Pictures and Make-believe." *Philosophical Review* lxxxii, 3, 283-319.
--(1978a). "Fearing Fictions." *The Journal of Philosophy*. 75: 5-27.
--(1978b). "How remote are fictional worlds from the real world?" *The Journal of Art and Art Criticism*, 37, 11-23.
--(1984). "Transparent pictures: On the nature of photographic realism." *Critical Inquiry*, 11, 246-277.
--(1990). *Mimesis as Make-Believe*. Cambridge, Mass.: Harvard University Press.
--(1993). Metaphor and prop-oriented make-believe. *European Journal of Philosophy* 1: 39-57.
--(1994). "Morals in Fiction and Fictional Morality, pt.1." *Proceedings of the Aristotelian Society, Supplementary Volumes*, 68: 27-50.

--(1997). “Spelunking, Simulation, and Slime.” In *Emotion and the Arts*. Mette Hjort and Sue Laver (eds.). New York: Oxford University Press.
--(2006). “On the (So-called) Puzzle of Imaginative Resistance.” In *The Architecture of the Imagination: new essays on pretence, possibility, and fiction.* Shaun Nichols (ed.). Oxford: Clarendon Press.
Wartenberg, T. E. (2007). *Thinking on Screen: Films as Philosophy*. New York: Routledge.
--(2006). “Beyond Mere Illustration: How Films Can Be Philosophy.” *The Journal of Aesthetics and Art Criticism*, Special Issue: Thinking through Cinema: Film as Philosophy, 64: 19-32.
Weinberg, J. & A. Meskin. (2006). “Puzzling over the Imagination: Philosophical Problems, Architectural Solutions.” In *The Architecture of the Imagination*, S. Nichols (ed.), Oxford: Oxford University Press, 175-202.
Weatherson, B. (2004). “Morality, Fiction and Possibility.” *Philosophers’ Imprint*, 4: 1-27.
Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: The MIT Press.
Wheatley, T. & J. Haidt (2005). “Hypnotic disgust makes moral judgments more severe.” *Psychological Science* 16: 780-784.
Whistler, J.M. (1890/1967). *The Gentle Art of Making Enemies*. New York: Dover.
Wiggins, D. (1987). *Needs, Values, Truth*. Oxford: Basil Blackwell Ltd.
Wilson, G.M. (1997). “*Le Grand Imagier* Steps Out: The Primitive Basis of Film Narration.” *Philosophical Topics* 25, 295-318.
--(2011). *Seeing fictions in film: The epistemology of movies*. Oxford: Oxford University Press.
Wimmer, H & J. Perner (1983). “Beliefs about beliefs: Representation and constraining function of wrong beleifs in young children’s understanding of deception.” *Cognition*, 13: 103-128.
Witzel, C., Valkova, H., Hansen, T., Gegenfurtner, K. (2011). “Object knowledge modulates color appearance.” *i-Perception 2*: 13-49.
Wollheim, R. (1980). Seeing-as, seeing-in, and pictorial representation. In *Art and its Objects: With Six Supplementary Essays*, 2nd Ed. Cambridge: Cambridge University Press.
Yablo, S. (2002). “Coulda, Woulda, Shoulda.” In T.S. Gendler & J. Hawthorne (eds) *Conceivability and Possibility*, Oxford: Oxford University Press, 441-492.
Zajonc, R.B. (1968). “Attitudinal effects of mere exposure.” *Journal of Personality and Social Psychology*, 9:1-27.
--(1984). “On the primacy of affect.” *American Psychologist*, 39: 117-123.