ABSTRACT

Title of Dissertation:	WHAT EMOTION DOES: AFFECT IN EMPATHY, ART, AND BEYOND
	Heather Adair, Doctor of Philosophy, 2020
Dissertation directed by:	Professor Peter Carruthers, Department of Philosophy

This dissertation puts forth a series of arguments about the role of affect in everyday cognition. I begin in chapter 1 by developing a generalized philosophical and scientific account of what "affective" states—a term encompassing emotions, moods, pleasures/pains, and felt desires—are and how they arise. From there, I address a number of debates in moral psychology, aesthetics, and philosophy of art that revolve around the function of affective states. In chapter two, I weigh in on a long-standing disagreement about the automaticity of empathy; I contend that different so-called "kinds" of empathy are not in fact automatic, and that an explanatorily robust model of empathy must account for the influence of affectively-laden "underlying values." In chapter three, I focus on the "processing fluency" view of aesthetic pleasure, which equates aesthetic pleasure with ease of perceptual processing. I critique and amend this view by highlighting the ways in which perceptual *disfluency* and *negative* affect also contribute *positively* to aesthetic appreciation. And, in chapter four, I attempt to redress

the so-called “paradox of fiction” by claiming that emotions do not require belief-states to be considered real and theoretically rational instances of emotion. To do this, I point to research on affective prospection and mind-wandering to argue that emotions must in principle be distinguished from our beliefs.

WHAT EMOTION DOES: AFFECT IN EMPATHY, ART, AND BEYOND

by

Heather Victoria Adair

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Peter Carruthers, Chair
Professor Jerrold Levinson
Professor Rachel Singpurwalla
Professor Dan Moller
Professor Elizabeth Redcay, Dean's Representative

Dedication

For my mom.

Acknowledgements

I am deeply indebted to the great many people who have supported me throughout my writing of this dissertation, my academic career, and more generally throughout my life as a whole. My greatest debt is to my mother, Vivyan Adair, who has been an unending source of unconditional love and encouragement. My mother first introduced me to the gripping world of philosophical thought when I was just 8 years old, and she has patiently and lovingly read every word that I have written ever since. My other great debt is to my supervisor, Peter Carruthers, who is a brilliant and prolific philosopher, an unparalleled advisor, and an incredibly kind and generous person. Working with Peter has been a privilege and an inspiration, and I cannot imagine having taken this journey without his incredible guidance. Many thanks are also owed to Jerrold Levinson, Rachel Singpurwalla, Susan Dwyer, Jill de Villiers, Peter de Villiers, and Todd Franklin, all of whom played important roles in my development as a scholar. While I have benefited tremendously from the help of many faculty members, colleagues, and friends over the years, few have pushed me or reassured me as insistently as Christian Tarsney and Jesse St. Charles, both of whom I am honored to call friends.

Table of Contents

CHAPTER 1

ALL ABOUT AFFECT

I. INTRODUCTION

People's lives are profoundly emotional. On a daily basis, we run a gamut of feelings both evolutionarily ancient and cognitively complex, from happiness, sadness, anger, and fear, to satisfaction, offence, pride, and awe. Indeed, on average, adults report spending approximately 90% of their waking days in one or more of these emotional states—and those are just the emotions that make it into conscious awareness![1] As scientific research on emotion has grown over the last several decades, it has become apparent that sentiments such as these are far from idle. Emotions impact how we represent and engage with others, make decisions, form judgements of value and preference, develop notions of moral right and wrong, and motivate us to take interest in the world around us. And yet, despite its intuitive significance, many questions remain about the precise role that emotion plays in everyday cognition.

The aim of this dissertation is to resolve some of those questions. In the chapters that follow, I address a number of long-standing debates in moral psychology, aesthetics, and philosophy of art, all of which revolve around the function of emotion. In chapter two, I focus on a well-established dispute over the

[1] These figures were found by polling participants at random intervals throughout the day and surveying them as to the presence of 18 emotional states that are relatively high-arousal, e.g. awe, anxiety, joy, etc. (Trampe, Quoidbach & Taquet, 2015). Given that low-arousal and mildly valenced emotions themselves provide appraisals about the environment (e.g. boredom, contentment), and given the plausibility of unconscious affect (see section III), we are arguably in *some* sort of emotional or affective state at all times.

automaticity of empathy and argue that a complete model of empathy must account for the underlying influence of emotionally-laden values. In chapter three, I amend a promising theory of aesthetic pleasure by highlighting the ways in which perceptual disfluency and negative emotions contribute positively to aesthetic appreciation. And, in chapter four, I attempt to redress the so-called "paradox of fiction" by claiming that emotions do not require belief-states to be considered real and theoretically rational instances of emotion. In short, each chapter is respectively intended to clarify emotion's functional roles in empathy, aesthetic preference, and engagement with fiction.

Since each part of this dissertation explores different aspects of emotional processing in different contexts, they may be read independently of one another and in any order. Taken together, however, these targeted investigations contribute to the broader goal of clarifying the nature of emotional appraisal in its own right. To set the tone for this overarching project, I begin here by laying out a highly generalized philosophical and scientific account of what affective states are (section II), how they arise (section III), and what sets them apart from one another (IV). These are themselves subjects of considerable debate, so delineating which metaphysical commitments we need to make before exploring affect's function is a necessary first step. Having established this general account, I will summarize the ways in which subsequent chapters aim to demystify the role of affect in empathy, art, and beyond.

II. WHAT AFFECTIVE STATES ARE: FEELING, MOTIVATION, & EVALUATION

Consider the case of Zoe the Zookeeper.

> Zoe has just accepted a job cleaning the lion enclosures at a local animal rehabilitation center. On her first day of work, Zoe wakes up with a knot in her stomach and is filled with a mixture of nervous excitement and anxiety. When she gets to the center, she is assured that the lions have all been put in an adjoining enclosure (separated by ten inches of unbreakable glass, no less) so that they cannot harm her. While tending the grounds near the glass barrier, a large, captive lion suddenly lunges ferociously in Zoe's direction. She shrieks and pulls away so quickly that she trips, landing painfully on the hard ground. Upon hearing the commotion, Zoe's new supervisor runs over to make sure everything is alright: Zoe is embarrassed by her dramatic reaction, turns red in the face, and laughs apologetically—of course she knew she was safe, why was she so frightened? Over the course of the day, the lions pounce ferociously at Zoe several more times, only now she is able to maintain her composure…she even finds their fruitless efforts somewhat amusing! At the end of her shift, Zoe leaves work satisfied with herself for a job well done.

In this vignette, Zoe is undoubtedly *feeling* quite a lot. She experiences anxiety, amusement, fear, pain, embarrassment, satisfaction, and more, all over the course of a single day. But, what exactly *are* these feelings? After all, at first glance states such as these appear to be remarkably different from one another across a multitude of dimensions. For instance, some of Zoe's feelings are negative or bad-seeming (e.g., the pain of falling), while others are positive or good-seeming (e.g., the pleasure of amusement). Some are about specific events or objects in the world (e.g. her fear of the lunging lion), while others do not seem to be about any one particular thing at all (e.g., her general sense of anxiety). Some are fleeting (e.g., her immediate surprise as the lion pounces), while others are likely to prove more

pervasive (e.g., her excitement throughout the day). Some involve marked bodily or behavioral changes (e.g., blushing in embarrassment), while others do not entail obvious physiological manifestations (e.g., calm satisfaction). And some involve simple, reflexive processing (e.g., terror at the sight of a large, predatory animal), while others appear to be predicated on more sophisticated cognitive appraisals (e.g., new-found confidence when the lions lunge yet again). How are we to make sense of such a disparate array of mental phenomena?

In everyday speech, we might—as I have done up until this point—characterize Zoe's first day on the job as one that is intensely *emotional* or full of *feelings*; and it certainly is! In scientific parlance, however, it would be more accurate to describe her day as one that is robustly colored by *affect*. For theorists working in the sciences and empirically-minded areas of philosophy, the term 'affect' encompasses a wide range of mental states: these include not only emotions, but moods, sensory pleasures and pains, and felt desires. Later in this chapter, there will be much to say about what makes these kinds of affect distinct from one another. Before we explore those differences, however, we should first lay out a picture of what seems to *unify* such prima facie disparate states under the same conceptual schema. What features do mental states like emotions, moods, pleasures and pains, and felt desires share in common?

Establishing this unifying characterization is admittedly a delicate task. Over the last century or so, scholars have increasingly turned their attention towards the underlying nature of affect, both as a singular category and with respect to particular affective states such as emotions. In that time, many debates have taken

root—most notably, debates about the *constitutive* components of affective states (see below), and the extent to which we might classify them as *natural kinds* (see section IV). These investigations have inspired several competing traditions, each of which emphasize unique defining criteria. Importantly, since the focus of this dissertation is on the *function* of affect, I do not wish to commit myself to any one of these metaphysical views here. Rather, it is my aim to motivate a generalized description of affect that appeals to some of the field's most widely converged upon insights. In so doing, I hope to make progress on the question of what affect *does* in ways that will be both fruitful and palatable to theorists across the board.

So, what shared insights have emerged from these debates? One of the more thoroughly explored points of contention in the affect literature centers on what Jesse Prinz has appropriately deemed the "problem of parts," or the problem of identifying the necessary and sufficient components of affective mental states—specifically, emotions (2004).[2] In the philosophical tradition, answers to this line of inquiry have varied tremendously since the time of the ancient Greeks.[3] In more recent years, however, three broad approaches to this "problem" have emerged as theoretical front-runners in both philosophical and scientific circles: those that argue for the constitutive role of (1) *phenomenology* and *physiology* in emotion (one's "feelings"), those that do the same for (2) *motivation,* and those that claim

[2] It is worth noting that much of the literature I will explore in this section focuses squarely on analyzing instances of emotion (rather than, say, moods or pains/pleasures), either because these writings pre-date contemporary research on the broader category of affect, or because they were specifically aimed at characterizing emotion as a unique mental phenomenon.
[3] Here, I will focus on approaches that have been elaborated upon in the last several decades, but early formulations of these views can be found in the works of ancient Greek and Medieval philosophers as well.

cognitive (3) *evaluation* as emotion's essential trait. Since each of these components will factor into our working model of affect, it will be useful to step back for a moment and review why an emphasis on the roles of *phenomenology*, *physiology, motivation,* and *evaluation* have both theoretical and intuitive appeal.

At the turn of the twentieth century, philosopher and psychologist William James placed special emphasis on the roles of (1) *phenomenology* and *physiology* with his articulation of emotional non-cognitivism. Non-cognitivism stands in contrast to what James described as the "common sense" model of emotion, wherein it is assumed that we perceive the world, then "cry, strike, or tremble, *because* we are sorry, angry, or fearful [emphasis mine]" (James, 1884). According to non-cognitivists like James, the causal arrows in this picture are amiss: we do not perceive, experience emotion, and *then* feel. Rather, he argues that emotions are constituted by awareness of automatic and reflexive (hence "non-cognitive") changes to our bodily states. On this view, when Zoe experiences e.g. fear, she does so because the mere sight of the lion triggers an immediate increase in heart rate, muscle tension, pupil dilation, perspiration, etc., and it is Zoe's awareness of those states that makes her afraid. Strip away the *feeling* of those unwitting bodily changes, the non-cognitivist contends (return the heart rate to normal, relax muscular tension, ease visual processing, calm perspiration, etc.), and there intuitively seems to be no fear left of which to speak.

While non-cognitivism nicely accounts for the apparent automaticity of our emotions and has been defended by prominent scholars (Robinson, 1995, 2004; Prinz, 2004; Zajonc, 1984), others have worried that this view seems to rob emotion

of its causal import. John Dewey, a contemporary of James, remarked that if someone like Zoe were really afraid *because* she e.g. experienced an increase in heart rate and turned to run, then her state of fear would play no explanatory role with respect to the pounding in her chest and her sudden startle backwards; fear itself would be causally inert (1895). These observations led Dewey to argue that specific emotional states are best identified by their (2) *motivational* contents. In modern versions of this tradition, emotions are often described either as evolutionarily adaptive mechanisms that shift one's action tendencies, or as states of action-readiness themselves (Deonna and Teroni, 2015; Ekman 1999; Tooby & Cosmides, 2008; Scarantino, 2015). According to these motivational theories, when Zoe is in a state of fear, she is in that state by virtue of her multifaceted disposition (either conscious or unconscious, occurrent or inactive) to flee, fight, or otherwise solve for the evolutionary problem of threat.

In the mid-20th century, a number of philosophers began to point out that views such as these neglect a crucial component of emotional processing, cognitive (3) *evaluation*. For, while it is clear that our responses to stimuli are often reflexive and fundamentally motivating, they also seem highly responsive to our complex *judgements* about the world. The first time the lion lunges at Zoe, for instance, she is presumably frightened because she evaluates her situation as one that is bad or threatening. But, as soon as that judgment changes (e.g. once she is reminded that she can observe the powerful animals without risk), she suddenly finds herself entertained by the lions' predatory displays. Philosophers in the evaluative or "cognitivist" camp have used this observation to identify emotions with the

tokening of particular judgments, i.e. to be afraid just is to judge something as dangerous, bad, undesirable, and so on (Nue, 2000; Solomon, 1988). This view nicely captures the intuition that emotions are fundamentally intentional, i.e. that they are object-directed or seem to be about things.

While these traditions have been highly influential and are still maintained by many, they are not exhaustive, nor need they be mutually exclusive. Indeed, many scholars are now starting to eschew the reduction of emotion to any one of these components. Instead, they opt for hybrid theories that claim constitutive roles for various combinations of physiological feeling, motivation, and evaluation (Helm, 2009; Tappolet, 2016). Roberts, for instance, suggests that emotion is best understood as a kind of perceptual experience of our "impressions" of the world, thereby attempting to account for both the sensory-like phenomenology of emotions and the fact that they seem to entail assessment (2003). Others have placed emphasis on identifying emotions as feelings, but have argued that such feelings are imbued with intentionality qua *feelings towards* things (Goldie, 2000).

Of course, different iterations of these views have different strengths and weaknesses, all of which are well worth considering in the context of the larger metaphysical debate.[4] For our purposes here, however, we need not make any such commitments about the necessary and sufficient conditions of emotions, nor any other affective state. By highlighting some of the considerations that motivated these traditions, my aim has merely been to show that prototypical instances of emotion reliably *have* such components, whether they play a *constitutive* role or

[4] Unfortunately, a full investigation of these nuanced views would take us far afield from the project at hand. For a more comprehensive review of these "traditions," see Scarantino, 2016.

not. When Zoe becomes afraid of the lion, (i) her heart-rate spikes and her breath quickens, (ii) many of these bodily changes are introspectively accessible upon reflection, i.e. if asked, Zoe would report that she felt afraid in that moment, (iii) she has a strong disposition to get away from the lion (so strong, in fact, that she stumbles to the ground while pulling back); and (iv) she instantly appraises the lion as potentially bad or threatening. Were any one of those reactions missing entirely, one might reasonably question whether Zoe was afraid at all.

Crucially, these components seem to obtain not only in cases of emotion, but in cases of mental phenomena like moods, sensory pleasures and pains, and felt desires as well. Though much separates these states conceptually (see section IV), they share this much in common. When Zoe wakes up with a *mood* of anxiety, for instance, that mood likely entails (i) heightened physiological arousal, (ii) feelings of unease (iii) motivation to be on the alert, and (iv) an evaluation that things are bad or could go very poorly. When she falls backwards and experiences *pain*, we can assume that (i) her blood pressure and cortisol levels spike, (ii) she finds the experience to be unpleasant, (iii) she is disposed to cry out and cradle her injured elbow, and (iv) she evaluates that sensation as negative. And, by the same token, were Zoe to *feel a desire* to e.g. eat a celebratory piece of chocolate at the end of the day, that felt desire might issue forth in (i) salivation at the sight of chocolate, (ii) awareness of her craving for a decadent treat (iii) taking steps to procure chocolate when able, and (iv) evaluating chocolate as a good or worthwhile thing.

This, then, is our bare-bones model of what affect is as a descriptive category. When scholars refer to affective states, they are pointing to mental states

that involve concerted changes in (i) physiology, (ii) phenomenology (ii) motivation, and (iv) evaluation. So far, so good. Unfortunately, simply identifying these important components does not tell us much about the *processes by which* affective states are reached—an issue that will be fundamental to exploring the role of affect in later chapters. To fill in this gap, we will have to look more closely at theories that deal with how our affective evaluations are formed.

III. How Affect Arises: Appraisal, Values, & Valence

In our brief overview of cognitivism, we saw that there are good reasons to emphasize the role of cognitive *evaluations* in our affective lives. After all, even when our responses to the world seem automatic, they also appear to be highly sensitive to context and framing, i.e. they require some sort of evaluation of events and objects in the world around us. But, how are we to understand evaluations themselves? As we saw earlier, some philosophers have claimed that evaluations are best understood as *judgments*. In early versions of the evaluative tradition, judgment is often implicitly characterized as an assent to some proposition, e.g., the judgement *that* I have been wronged (anger), *that* I am in danger (fear), *that* a situation is beneficial (happiness), etc. But, equating all of our emotional, moody, painful, etc. states with mere assent to propositions turns out to be rather problematic.

Most critiques of judgmentalism turn on the fact that one can consciously assent to a proposition without instantiating any of the other components that we have come to associate with affect. One could, for instance, make the judgement

that "so-and-so has insulted me" without being *motivated* to do anything about it. If one thinks that said insult is in no way important or germane to one's daily life, that judgment might remain motivationally inert with respect to both disposition and action.[5] Relatedly, merely assenting to the proposition that one has been insulted does not necessarily result in the *physiological* and *phenomenological* changes we associate with anger, i.e. one can acknowledge that they have been insulted without *feeling* angry. And most worrisome, judgmentalist formulations of evaluation struggle to account for "recalcitrance" sentiments, or emotions that seem to persist *despite* the presence of judgments that should undermine them (D'Arms & Jacobson 2003). Zoe could, for instance, find herself feeling frightened by the lions on her second day of work, even though she now genuinely avows that she is not in danger.

Clearly, then, the evaluations that undergird our affective states cannot be judgements *simpliciter*. They must be *special kinds* of judgments that link up predictably with all of the elements listed in our generalized model of affect. But, of course, this observation tells us little about how those special judgments arise or why they coincide with changes in feeling and motivation. Fortunately, in the last sixty years, cognitive psychologists have made considerable progress on this issue by elaborating on the concepts of *appraisal* and *valence*.

Simply put, affective appraisal is the process through which an individual determines the subjective significance of any given situation, event, or object

[5] If, for instance, you were traveling in a foreign country where some gesture were considered the height of insult, you could feasibly be on the receiving end of that gesture and judge *that you have been insulted* without feeling any motivational force to do anything about it or even, for that matter, *care*. For more on the relationship between affect and caring, see chapter 2.

(Arnold, 1960). When Zoe feels afraid of the lion, that sentiment seems to be predicated on an *appraisal of the lion* as something that is, for instance, bad. Crucially, what makes this act of affective appraisal distinct from generic judgment is the process by which it occurs. While a traditional characterization of judgment hinges on one's higher-order assent to propositional content, most models of appraisal allow for evaluative pathways that are both (1) unconscious, fast, automatic, associative, stereotypic, etc., and (2) conscious, slow, effortful, logical, etc. (see e.g. Kirby & Smith, 2000, Scherer 2001). These pathways correspond roughly with Daniel Kahneman's distinction between "system 1" and "system 2 processing," and as such will be differentially recruited under specific conditions, i.e. stress, fatigue, type of stimuli, etc. (2011).

This process model of appraisal has considerable explanatory power. First, it elaborates on the sense in which even knee-jerk, "automatic" emotions involve something like evaluation. When Zoe is first startled by the lion, it is because she immediately appraises the lunging animal as bad via reflexive, associatively learned or innate mechanisms,[6] hence her instantaneous and violent attempt to flee. Soon, however, Zoe's more complex appraisals of e.g. her relative safety, the importance of professionalism, etc., dominate, and her fear is quelled. In short, both responses involve appraisals, they are merely generated through different pathways. Crucially, although these appraisal processes are independent, they are not isolated. One's overall affective state is not the result of a single instant of appraisal: rather, our affect shifts and changes over time through extensive and

[6] Studies have shown that humans are predisposed to automatic fear responses in the presence of e.g. large, predatory animals (Garcia, 2017)

continuous *re*-appraisal. Unconscious appraisal of another person as dangerous might, for instance, output heightened arousal and motivational tendencies to guard oneself against attack—those physiological/motivational *outputs* can then be taken as *input* to more conspicuous appraisals of that person as untrustworthy or threatening. This is why Zoe's fear may abate gradually, and in ways that are responsive to changes in her environment and internal states.

Second, this model explains what judgementalism could not with respect to recalcitrant emotions. More often than not, our reflexive and reflective assessments of the world inform and reinforce each other.[7] That said, because these two types of processing are especially sensitive to different types of information, we would also expect them to dissociate from time to time. In particular, associative processing is sensitive to perceptual information that may not be semantically encoded (and thus is largely unavailable to conscious reasoning). Reasoning, by contrast, allows us to e.g. draw connections between one's current and prior experiences, make logical inferences, etc., all of which may be unavailable for associative processing until they have been made and stored in memory. So, although Zoe may consciously appraise herself to be safe, an unwitting appraisal of the lions' sharp teeth and powerful claws may override those reflective judgments in determining that stimuli's significance—namely, as e.g. bad or imminently harmful.

[7] For instance, you might consciously appraise your friend to be a good person because of the generous ways that she has treated you over the years, while also unconsciously appraising her familiar demeanor and non-threatening body language as pro-social. Both of these appraisals would reinforce your overall assessment that she is, in fact, a lovely person.

It is worth noting that an object or event's significance, in this sense, cannot be appraised in a vacuum. Indeed, how Zoe evaluates the world seems to be very much a function of things like her goals, intentions, and values. Consider Zoe's embarrassment after falling in front of her new supervisor. Assuming that Zoe *values* being seen as a competent employee, she would likely appraise this interaction as negative. But, suppose for a moment that Zoe did not in fact want to keep her job at the zoo; she took the position as a summer job to placate her parents, and clandestinely hopes to be fired as soon as possible. Under those conditions, Zoe would likely appraise her employer's looks of concern as significant in a *positive light*—she might even ham up her performance as a clumsy and incapable worker! Interestingly, this dependence on one's underlying values is evident even in the case of appraisals that are swift and automatic. For, although human beings seem to have an innate fear of stimuli that are snake-like *(*Roach, 2001), a herpetologist with years of experience handling snakes safely would likely not experience immediate fear in the presence of a harmless, coiled reptile.[8]

Appraisal, so construed, is a context-sensitive process that determines significance. Appraisals of different *kinds* of significance appear to directly result in changes to one's physiology, motivation, and phenomenology. An appraisal of dangerousness might, for instance, activate one's fight-or-flight response. In this state, one's heart rate and breathing become rapid, peripheral blood vessels in the skin constrict (which is why we grow "pale with fear"), attention becomes narrowed

[8] Though I will not explore the issue at length here, in chapter 2 we will see that there has been some confusion over whether or not particular *sorts of concerns* give rise to different *kinds* of valence. I will try to clarify this issue there.

in on the source of danger and potential routes of escape, and muscles tense in preparation for e.g. flight. All of these physiological and motivational changes impact not just how we behave, but how we feel, i.e. they can be accessible in conscious awareness and often contribute to what it feels like to be e.g. in a panic. But, importantly, these are not the only things that impact the phenomenology of affective states. Appraisal also generates valence.

Valence is, very roughly speaking, what makes our affective states seem good or bad, pleasant or unpleasant, attractive or aversive, etc. While there is some disagreement over which of these characterizations is more apt, there is strong consensus that valence is what makes our emotions, moods, pleasures and pains, and felt desires either positive or negative (see Frijda, 1986; Ortony et al., 1990; Solomon and Stone, 2002; Russell, 2003; Charland, 2005; Colombetti, 2005; Barrett, 2006). Going back to our working example, Zoe's anxiety, fear, pain, and embarrassment are negatively valenced, while her excitement, satisfaction and amusement are positive.

Importantly, this demarcation between negative and positive valence is not merely conceptual. The hedonic tone of our appraisals *in general* seem to be encoded within a number of limbic areas including the anterior insula, the ventromedial prefrontal cortex and the orbitofrontal cortex (Rainville et al., 1997; Peyron et al., 2000; Levy & Glimcher, 2012). Furthermore, there is good evidence that unique mechanisms exist for the processing of e.g. pleasant and unpleasant stimuli (Lewis et al., 2007; Viinikainen et al., 2010). Specifically, scientists have found regions of the brain that correlate exclusively with negative valence (e.g., the

dorsolateral prefrontal cortex, anterior midcingulate cortex, frontal pole, inferior parietal cortex), and others that correlate with positive valence (e.g., substantia nigra, ventral striatum, right caudate nucleus; see Colibazzi et al., 2010). Granted, this binary model of valence does not mean that all affective experiences must be unambiguously experienced as positive or negative in any given instant. One could, for example, realize that she has been insulted, appraise that state of affairs as negative, and feel angry, while also appraising her own moral outrage as an appropriate, good-seeming and hence pleasurable response: hence, righteous indignation. Even in the case of "mixed emotions," however, changes in regions of the brain that signal valence seem to correspond directly with the mélange of good and bad-seeming sentiments that we experience. [9]

As an output of the appraisal system, valence has a number of important features that account for its relation to the motivational, physiological, and phenomenal components of affect. As I previously noted, valence appears to be causally linked to the phenomenology of affect, i.e. the *sensations* of good and bad-seemingness that we experience when in affective states. Perhaps unsurprisingly, neuroimaging studies suggest that clinically depressed patients tend to have lower activation of positive valence systems during reward tasks and greater baseline negative valence activation when compared to nondepressed controls (J. P. Hamilton et al., 2012; Pizzagalli, 2014). These differences account for depression's ability to rob individuals of the phenomenal pleasure we associate with more

[9] Whether we are, in these cases, flipping back and forth between valenced appraisals or simultaneously tokening contradictory valenced appraisals is an interesting and open empirical question that I will address in chapter 2.

positive emotional states. Furthermore, valence appears to be the mechanism by which states like pain and pleasure derive their phenomenal content. For example, notice that pain is not *merely* a representation of one's bodily condition. Although people on morphine are able to accurately report the sensory-motor qualities of their injuries, they no longer feel those sensations *as bad*. Even on morphine, Zoe would still experience a descriptively sharp, stabbing sensation in her right elbow upon falling, she simply would no longer feel it *as* painful. In short, when valence is mitigated, it directly impacts what we *feel* in conscious awareness.[10]

It is also generally assumed that valence is intrinsically motivating. In both philosophy and the sciences, there has been a long tradition of viewing e.g. pleasure and pain as fundamentally motivational. When one e.g. feels pleasure while eating chocolate, that pleasure presumably functions as a signal that *eating chocolate is good (*or that *chocolate is good),* which serves as a catalyst for action. But, notice in this case that it is not the overall *feeling of pleasure* that is motivating, but *valence as an indication of seeming-goodness*.[11] This distinction is important, since it allows us to make sense of the fact that valence plays a fundamental role in *both* conscious and unconscious decision-making. It has been

[10] Similarly, sensorially pleasurable experiences are not reducible to mere sensory representation. One may derive great pleasure from eating chocolate, but eating *tremendous, unending* quantities of chocolate over time will soon cease to be pleasurable. This is because satiety mitigates positive valence, even though the gustatory qualities of the chocolate remain constant.

[11] It is very worth noting that seeming-goodness and pleasure are not identical. One might feel considerable sensory pain when receiving a flu shot (because the prick of the needle is appraised as bad-seeming), but still affectively appraise that medical procedure *as good* overall. In fact, this appraisal of painful activity as good-seeming has clinically been shown to decrease overall negative valence, and hence mitigate the painfulness of pain. As such, valence qua seeming-goodness/badness seems to be what motivates us to action. For a more thorough defense of this characterization of valence, see (Carruthers, 2018).

shown, for instance, that when we prospect—when we imagine various options or courses of action available to us—valence-signals are ultimately what determine the choices that we make (Gilbert & Wilson, 2005). That is to say, it is the appraised positivity and negativity of our options that leads us to act as we do. Conscious prospection *may* often include phenomenally conscious bouts of pleasure that are motivational, as when prospection about vacationing in Spain *feels pleasurable*. But, neither prospection nor valence need to be consciously accessible. Much of our rapid-fire, affectively laden decision-making occurs outside of conscious awareness (e.g., when you select one chair out of many to sit in), and even conscious acts of prospection need not produce *felt valence* to result in motivation.[12]

The fact that valence plays this strong motivational role has noteworthy theoretical ramifications. It has led many to suggest, for instance, that valence functions as a singular, "common currency" for comparing, ranking, and choosing between diverse options. (e.g., McNamara and Houston, 1986; Cabanac, 1992; Montague and Berns, 2002; Pfister and Böhm, 2008; Levy & Glimcher, 2012). On the common currency model, in order to choose between two very different alternatives—say, going to the gym vs. staying home to indulge in that piece of chocolate—one would have to be able to compare those options *on a single dimension* (as I have articulated it so far, seeming-goodness and seeming-badness). That is because positing multiple *types* of positive and negative valence makes these activities look rather incommensurate. If there were one type of

[12] Both appraisals and valence, then, can arise outside of conscious awareness.

valence for e.g. goal-congruence and another for pleasure-conduciveness, it isn't clear which activity would seem like *the better* option (gyms are good for goals, and chocolate is good for pleasure). If valence is the summation of a target's overall positive and negative value, however, we would be able to weigh valence-signals and come to a concrete decision.

Whether or not valence is *itself* a single natural kind, and whether it contributes to affective states in ways that make *them* natural kinds is a highly contentious topic. "Core affect" theories maintain that valence is one-dimensional. They argue that affect as a broad category is a natural kind because it varies reliably along stable axes of valence and arousal while denying that status to our folk-concepts of unique emotions (Russell, 1980; Russell & Barrett, 1999; Barrett & Wager, 2006). These are also often referred to as "psychological constructionist" views, because they claim that emotions are the result of downstream social construction and internalization. This stands in opposition to "categorical" theories of affect, which place emphasis on the notion that we have innate, discrete operating systems for each 'basic' emotion.[13] On this account, some of our affective folk-concepts do pick out natural kinds—specifically, those that are individuated by different *types of valence* and patterns in neurobiological, physiological, expressive, behavioral, and phenomenological responses (e.g. Ekman, 1992; Panskepp 1998). Still others maintain that we have some "affect program emotions" that may be classified as natural kinds, while others are higher-order and socially constructed (Griffiths, 1997).

[13] For more on the debate between categorical and dimensional models of affect, see e.g. Zachar & Ellis (2012).

In chapter 2, I will spend some time defending a one-dimensional characterization of valence that should be amenable to both "categorical" and "dimensional" sides of this debate. The question of natural kinds is not, however, one that I will be addressing here. Whether or not our folk concepts of emotion constitute natural kinds is currently an open and worthwhile question. So too is the status of affect as a broad category. But those open questions should not have any impact on our ability to theorize about the *function* of states which bare the affective markers we have explored thus far.

IV. KINDS OF AFFECT: TARGETS & CAUSES

At this point, we have seen that affective states entail concerted shifts in physiology, phenomenology, motivation, and evaluation. We have also explored the roles of appraisal and valence in eliciting those changes. But, if all of our emotions, moods, sensory pleasure and pains, and felt desires share these elements in common, what sets them conceptually apart?[14] To address that question, let us first establish some terminology that will be useful both here and throughout this dissertation.

By and large, phenomena such as emotions, pleasures and pains, and moods are distinguished by the nature of their "targets." The target of an affective state is simply the object at which an appraisal of goodness or badness is directed; it is what the feeling is *about*. Because affective appraisals are always elicited *by* something,

[14] All of the characterizations that I provide here are aimed at laying out broad, descriptive definitions of emotions, pleasures and pains, moods, and felt desires—I hope to have done so in a way that is roughly compatible with ordinary linguistic usage. As before, I do not intend to make any metaphysical claims about whether these boundaries obtain in producing natural kinds; indeed, many folk-concepts of affect seem to span these categories or altogether defy categorization.

i.e. are reactions *to* something, it is unsurprising that they are generally *about* something in this way. But, notice that an affective state's represented *target* is not necessarily what *causes* that appraisal in the first place. Over the last few decades, many researchers have found that appraisals of seeming-goodness and badness can be caused by one stimulus and then misattributed unwittingly to unrelated targets. Reading moral vignettes while in a dirty room, for instance, has been shown to intensify sentiments of moral outrage and anger towards the contents of the vignettes themselves. While the target of these subjects' anger is the moral vignettes, the cause is presumably the participant's reaction to filthy reading conditions (Schnall et al, 2008). This effect has been shown to be surprisingly pervasive under the right conditions: it has been observed in relation to ratings of attractiveness in the presence of spirited music (Marin et al, 2017), romantic interest in the presence of danger (Schachter & Singer,1962), social preferences in the presence of particular scents (Li et al, 2007), and gustatory preferences in the presence of smiling vs. frowning faces (Winkielman et al, 2005).

To see why this misattribution is possible, recall that our affective appraisals can and often are reached via unconscious, associative processing (see section III). If the cause of an affective appraisal is subtle (largely perceptual, nonobvious, or hidden from focal attention) it can produce valence to which we have a hard time attributing a target. In the study by Winkielman et al., for instance, affect was only misattributed when its causes were inaccessible to conscious awareness. In that study, the presence of frowning and smiling faces only had an impact on beverage preference ratings when those faces were backwards masked, i.e. presented

subliminally. Similarly, Li et al. found that smells could shift social preferences only when those smells were barely detectable. In these cases, with little environmental information to go off of, participants likely inferred through conscious reasoning that the valence-signal they felt was related to more "obvious" but incorrect stimuli.

So, how do causes and targets bear on our ability to differentiate between kinds of affect? Well, to begin with, all affective states have causes and targets of one sort or another (they are all evoked, and they are all about something). In the case of emotion, however, those targets tend to be clearly specified or delineated, i.e. they are directed either at particular *entities* or *states of affairs,* and as such can usually be expressed by a proposition (de Sousa, 1987). To use our working example, Zoe is afraid *of* the lion, she is satisfied *that* she had a successful first day, she is embarrassed *about* her dramatic reaction, and so on. This object-directedness is taken by many to be a defining characteristic of emotion. In our scenario, Zoe's fear, satisfaction, and embarrassment are all considered instances of emotion because they are directed at concrete targets such as the lion, success, and social opprobrium, respectively. The targets of these and other emotions tend to follow their eliciting stimuli closely or even instantaneously. Although emotional misattribution is certainly possible (see above), the cause of one's e.g., fear and the target of that fear are more often than not one and the same, i.e. if a lion's powerful lunge causes negative appraisals, the lion will most likely become the target of that appraisal.

But, not all affective states take on such clear and causally potent targets. Unlike emotions, moods are often distinguished by the targeting of more diffuse and nebulous objects (Frijda, 2009). When Zoe wakes up in the morning with a knot in her stomach, for instance, it is not necessarily obvious (either to herself or to others) what, precisely, her state of anxiety is *about*. Her nervousness might be directed at something as general as her *entire* life trajectory, her sense of self-worth as an individual, or the state of the world at large: it could be about anything! This vagueness may arise for a number of reasons. For instance, it is possible that the initial cause of Zoe's appraisals was itself vague, e.g. she really is appraising her overall self-worth—a thing for which she has no clear conceptual boundaries. It is also possible that her mood did have a concrete cause, but that for whatever reason the resulting appraisal has been misattributed to an ill-defined, non-matching target (Morris, 1992). That is to say, the cause of Zoe's negative mood may very well be discrete—she may feel nervous *because* she ill-advisedly watched *When Zoo Animals Attack* the night before her first shift—but fail to attribute those bad feelings to the documentary itself; as such, she is left with no obvious target for her representations of seeming-badness and a general mood of anxiety.

Given these broad characterizations of emotion and mood, it might seem as though we have covered all instances of affect by default. After all, an affective state either takes a target that is representationally identifiable or it does not.[15] So, where does that leave us with pain and pleasure or felt desire?

[15] This is not intended to sound tautological: the underlying appraisals that drive a mood can (and often do) go on to drive an emotion when those appraisals are attributed to a clear target; conversely, an emotion can become a mood when its target becomes conceptually confused or nebulous, e.g. general anxiety can become fear, and fear can produce general anxiety.

On reflection, it would appear as though pain and pleasure take on targets that are rather precise and clearly delineated, making them seem more like *special* instances of emotion than anything else. In these cases, we conceptually draw the boundary around emotional states whose targets specifically involve one's bodily sensations (touch, taste, sight, sound, smell, etc.).[16] When Zoe startles and lands squarely on the ground, the target of her mental state is presumably the stabbing *sensation* produced in, say, her right elbow upon impact. Her 'pain' is constituted by the seeming badness of *that* sensation. Granted, the *occurrence* of a sensation can also be the target of non-sensory emotional appraisal; were Zoe's affective state to take the *descriptive fact of her bruised elbow* as its target (rather than the sensation itself), she would likely be sad, angry, embarrassed, etc. about the injury. But, notice that while we might metaphorically proclaim that she is "pained" in this scenario, the object of those particular emotions is not itself a sensory representation.

This sensory vs. non-sensory distinction is not always easy to maintain, especially in light of affective misattribution and the impact of top-down expectations such as the placebo effect. For example, the bad-seemingness tokened while feeling sad about the *occurrence* of the injury may very well be misattributed to the sensation itself, thereby making Zoe's sensory pain worse than it would have been otherwise. Conversely, finding her own fall humorous or entertaining could dampen her subjective feeling of pain as she experiences it. This phenomenon has

[16] I specify "sensory" pleasure here to eschew its vernacular (often metaphorical) use which encompasses arguably non-sensory "pleasures" such as joy and happiness; by the same token, "sensory" pain is intended to exclude non-sensory instances of "suffering" such as depression and grief.

been observed in a number of contexts, most notably when higher-order expectations of perceived benefit mitigates subjective experiences of sensory pain. For instance, when we expect that an injection will bring long-term protection from e.g. the flu, we tend to appraise the sensation of receiving that injection as less bad-seeming than we do when the prick of the needle serves no apparent purpose.[17] Nonetheless, even in these sorts of examples, one can still in principle distinguish between emotions that take sensory and non-sensory representations as their primary targets.

Finally, felt desires are unique insofar as they seem to be distinguished not by the nature of their targets, but by the presence of an especially strong motivational component *towards* those targets. To be clear, not all desires are felt or occurrent, i.e., not all desires operate in conscious awareness at all times. When Zoe wakes up in the morning, she likely has a vast number of *standing* desires that are not psychologically active while crawling out of bed. She might desire world peace, to meet the love of her life, or to have a car with better gas-mileage, all without feeling those desires in the moment. When a desire is felt, however, it is accompanied by the instantiation of action tendencies that will, *ceteris paribus,* bring about her desired state of affairs: all things being equal, desiring chocolate will result in Zoe attempting to procure chocolate. Whether her motivational tendencies are actually acted upon, though, will depend not only on Zoe's desire for chocolate, but on her other (potentially conflicting) desires, beliefs, goals,

[17] Whether this is changing the *sensory-motor* representation of the painful experience (hence eliciting a "pure" affective response to the new sensory representation) or is impacting the emotional appraisal itself is a question of some debate. See Ellingsen et al, 2013.

intentions, and expectations—being primed to obtain chocolate might not result in action if Zoe has a strong competing desire to avoid sugar or believes that chocolate will be deleterious to her health.

Notice that because felt desires can take on any target, we should expect them at times to be classified in ordinary language-use as *kinds* of emotions, and at other times as *kinds* of moods. When the target of one's motivated appraisal is apparent (when one wants *that piece of chocolate*), that desire might be described in emotional terms like "lust" or "yearning:" Zoe is feeling lustful towards the chocolate because the chocolate is good-seeming to her and she is now primed to get it. But felt desires can also be mood-like, insofar as they sometimes take on nebulous or unclear targets. Early pangs of hunger or thirst are prime examples of this, since one may feel motivation to find *something* to slake one's thirst or hunger without know what, exactly, that thing might be (hence the tendency to stand in front of one's refrigerator prospecting about different possible targets). Similarly, so-called moods of "restlessness" and "ennui" are often accompanied by motivations to do *something*, without a clear targeting in mind.[18]

V. DEBATES AND PUZZLES

[18] In recent years it has become apparent that distinct subcortical networks correspond with everyday use of the word "desire." One is "liking," which involves phenomenally accessible positive valence (e.g. that chocolate looks good); the other is "wanting" or "having an urge," which is associated with action tendencies and can dissociate from positive valence (Berridge & Kringelbach, 2008, 2013). For instance, an addict might want a potent, life-ruining drug, and feel considerable displeasure when acting upon the motivational tendencies that wanting drugs entails.

This dissertation is, as the title of the chapter would indicate, all about affect. And now, we have a general (hopefully uncontroversial) working characterization of what affect is, how it comes about, and what makes our folk-concepts of affect distinct in theory. From here, we will focus on what affect *does.* Over the years, a number of interesting puzzles and debates have cropped up around how the affective system interacts with and impacts other cognitive capacities. These debates include questions about the role of affect in learning, memory, decision making, perception, moral judgement, motivation, reasoning, biases, aesthetic judgment, moral responsibility—the list goes on and on. In the chapters that follow, I set my sights on the role of affect in three domains: the generation of empathy, aesthetic appreciation, and (relatedly) engagement with works of fiction.

We experience a broad range of emotions in response to other living creatures, at times with great empathy. In Chapter 2, I will argue that while the ability to "feel as others feel" is clearly of ethical import, our current models of empathy are incomplete. To date, most theorists have focused on two cognitive systems as the foundations of empathy: emotional mirroring and perspective-taking. While investigations into these mechanisms answer important questions about how we engage with others, they largely ignore the affectively-laden *underlying values* that make us care about those others in the first place. Recent research into the effect of group status on emotional contagion and work on subjects with e.g. psychopathy, and sadism indicate that mirroring and perspective-taking is not (as some have implied) sufficient for explaining when empathy does and does not arise. Rather, these systems appear to rely initially on the *valuing* of others and their emotional

states/well-being to elicit full-blown empathy. Because values have been relatively overlooked in the literature, there are many pressing questions regarding their status as discrete mental states, how conflicting values are affectively represented, and to what extent we are morally responsible for their formation. By exploring a few of these open questions, I hope to lay the groundwork for a more comprehensive model of empathy.

Unsurprisingly, our understanding of affect also has strong implications for the field of empirical aesthetics. Over the past several decades, researchers have become increasingly interested in explaining judgments of beauty. The "processing-fluency" view, for example, identifies aesthetic pleasure with the positive affect we feel when we are able to easily process an object's perceptual properties. In chapter 3, I critique this theory and endeavor to highlight what it neglects: the importance of perceptual *disfluency* and *negative* affect. By turning to contemporary work on the affective mechanisms driving e.g. attention, curiosity, and problem-solving, I show that disfluency and the negative affect that it evokes are often crucial for motivating, amplifying, and sustaining ongoing aesthetic pleasure. I go on to argue that these elements often trigger primitive "questioning attitudes" such as interest, which are fundamental for genuine aesthetic appreciation. I have found this conception of interest to be particularly fruitful and am keen to delve further into the topic across a number of domains (see below).

Finally, theoretical models of affect have had a profound impact on topics in aesthetics and philosophy of art. The "paradox of fiction," for instance, is built upon the cognitivist premise that emotions require existence-beliefs, i.e. that to be truly

afraid of, say, a bear, one must *believe* that said bear poses an extant threat. Since engaging with fiction does not seem to entail such existence-beliefs, many philosophers have argued that our apparently emotional responses to fiction are paradoxical. In chapter 4, I address this paradox by arguing that affective states do not require judgements of truth (i.e. belief) in order to count as instances of real and/or theoretically rational emotion. I draw upon empirical findings on the recruitment of affect in counterfactual mind-wandering and prospection to demonstrate that the affective system routinely appraises the seeming goodness and badness of imagined scenarios, even when we know those scenarios are not real.

Many of the themes that I have tackled in these chapters while exploring aesthetics and moral psychology also have broader ramifications in philosophy of mind. Moving forward, I look forward to continuing my research on the ways in which affect drives understanding. In philosophy of science, understanding is often associated with the application of laws of logic, rules of inference, principles of causality, etc., and the "coherence" that these provide to make understanding possible. But, relatively few theorists have considered the extent to which emotional appraisals both spur our efforts to understand, and serve as internal indicators of the success or failure of those explanatory models. Given my current work on the motivational role of affect, the significance of underlying values, and recent findings on information-gathering tendencies in human and non-human animals, I want to develop the claim that affective states deeply impact our ability to learn, solve problems, and exercise willpower. This work on understanding should also challenge and amend leading theories of general intelligence and

willpower—which often focus primarily on the impact of working memory and executive function, respectively—by showing that these capacities are moderated by how affectively interesting and valuable one finds a particular task (or tasks in general) to be. This thesis should be empirically testable, and if plausible will have clear, important pedagogical applications.

CHAPTER 2

UNDERLYING VALUES:
THE ROLE OF APPRAISAL IN EMPATHY

I. INTRODUCTION

It is standard practice across the disciplines to hedge any discussion of empathy with a caveat: empathy is difficult to define. Because our everyday folk conception of empathy captures a wide range of cognitive capacities, theorists have developed two broad research programs to study the phenomenon. Specifically, they have set their sights on exploring two different *kinds* of so-called empathy: (1) "affective empathy" and (2) "cognitive empathy." As the name might indicate, (1) "affective empathy" refers to the mere tokening of *shared affect*. On this characterization, to empathize one must simply instantiate another person's emotional state first-personally, i.e., one must feel that emotion oneself. By contrast, (2) cognitive empathy is characterized as an instance of *vicariously shared affect*. This requires not only that one feel another's emotion, but that one represent and understand the other person's experience *as such*. In general, then, empathy is commonly assumed to have two primary components: an affective component (the ability to share the emotions of another), and a cognitive component (the ability to understand the perspective of another).

The schemas proposed above have, no doubt, proven empirically fruitful. Over the years, countless empirical studies and theoretical advancements have improved our grasp of the mechanisms driving (1) shared affect (affective empathy) and (2) vicariously shared affect (cognitive empathy), respectively. Affective

empathy, for instance, is often equated with neurological mirroring, emotional contagion, and other forms of "low-level" emotional recognition. Research in this domain has taught us much about the complex ways in which human beings unwittingly detect, experience, and respond to others' emotional expressions and behavioral cues. At the same time, our understanding of cognitive empathy has been enriched by both scientific and philosophical work on "mindreading," or the ability to attribute mental states to others. Since cognitive empathy requires that we feel others' emotions *in light of* those others' e.g. beliefs, goals, desires, etc., establishing how those representations are tokened is an important step in demystifying the nature of vicarious affective sharing.

To date, theorists interested in explaining empathy have largely focused their attention on these two domains. And that is a good thing, since mirroring and mindreading clearly contribute to the sharing of affect (vicarious or otherwise)! But, while these mechanisms may very well be crucial for empathy, *they do not fully explain it*. As it stands, most models of empathy highlight the ways in which mirroring and mindreading lead to empathy as stand-alone processes. It is important to note, however, that mirroring and mindreading do not arise in a bubble. As we will see throughout this chapter, people are not always prone to (1) "catch" others' feelings, nor do they reliably (2) take others' perspectives. If these capacities are fundamental for empathy—as seems to be the case—then a robust explanatory framework ought to address *why* they are differentially instantiated. *Nor do these mechanisms, on their own, suffice for empathy,* whether affective or cognitive. They need not only to be triggered, but maintained. What, then, is missing in this picture?

For a more complete structural model of empathy, I propose that we must turn our attention to the ways in which human beings both assess and value the world around them. To that end, I argue that in order to fully account for how empathy arises, we must integrate our understanding of the complex affective appraisal system that drives our emotional reactions. In section II, I will briefly review the literature on mirroring and mindreading to show how these systems work, and what they do and do not contribute to empathy. In section III, I describe how the affective-appraisal system seem to function, with special emphasis on what I will call "underlying values" and valence. Along the way, I aim to resolve some theoretical disagreement over the nature of valence by arguing that it represents a single axis of seeming-goodness and seeming-badness. With these pieces in place, in section (IV) I use the process of affective appraisal to explain both pathological and everyday failures to empathize—an explanation we cannot provide, I will show, without appealing to the appraisal system. In the hopes of motivating further investigation, conclude in section IV by raising a number of open questions about how our underlying values are set.

II. SHARED AFFECT & VICARIOUSLY SHARED AFFECT

As I mentioned from the outset, there is little doubt that mirroring and mindreading play important roles in the generation of empathy. But, of course, they do not *just* contribute to empathy: both are complex cognitive processes in and of themselves. Before we examine the ways in which our *underlying values* impact these systems through affective appraisal (see section IV), we should first briefly review what

they entail. Specifically, we should look to both empirical and philosophical research on the unique ways in which mirroring and perspective-taking are thought to contribute to our experience of others.

A. MIRRORING

Let's start with the phenomena of "mirroring." For several decades now, it has been noted that observing another's action and performing that same action can produce strikingly similar patterns of neurological activity (Rizzolatti & Craighero, 2004; Keysers, 2010). When an observer witnesses someone filling a coffee cup, for instance, many regions of the observer's brain will activate *as if* she were pouring the coffee herself. Neurons that are equally responsive while observing an action and performing it are often referred to as "mirror neurons."

Early empirical work on this phenomenon suggests that primates have a strong tendency to neurologically mirror goal-directed activities, i.e. grasping, holding, reaching, etc. (di Pellegrino et al, 1992). More recently, fMRI have revealed these "mirror-like" neurological responses in human beings, especially in the premotor cortex, the supplementary motor area, and the primary somatosensory cortex (Molenberghs, 2009; for review see Rizzolatti & Craighero, 2004). These regions of the brain are associated with spatial and sensory guidance of movement, the planning of actions, and proprioception, i.e. representations of one's own body. Because of this mapping, it has been suggested by some that mirror neuron activation might help us apprehend the intentions and goals of others. On this view,

offline activation of the motor-plans we observe in others enables us to represent and accurately predict how those others will behave (Iacoboni et al, 2005).

Although it is not clear that action mirroring does in fact give rise to these insights,[1] it does likely function as a priming function for first-person action (Byrne, 1998). Most research on this topic emphasizes the idea that mirroring occurs outside of conscious awareness. For our purposes here, however, it is important to point out that mirroring is also often accessible to consciousness awareness; when this occurs, it may indeed provide significant insight into *what it is like* to be in another's shoes. Think, for instance, of observing someone else as they perform a particularly high-stakes or difficult action, e.g. watching a friend as she tries to extract a poorly placed wooden block in a game of Jenga (one that is particularly difficult to extricate). When fully attentive in such a situation, one is likely to not merely exhibit neurological mirroring and activate similar action-plans offline, but to *feel* that mirroring in one's hands, to fight an occurrent *impulse* to extract the block oneself, and to otherwise have an embodied sense of what it would be like to be performing that difficult task. In other words, offline and unconscious mirroring can quickly come conscious. This is particularly salient when observing those who fail to properly execute a goal-oriented action. For example, while intently watching someone try to e.g. stay vertical on a balance beam, one may quickly *feel*

[1] Many scholars have suggested that this kind of mirroring does not actually provide us with information about others' mental states. Rather, they suggest, this neurological activity may arise as a repercussion of having already represented others e.g. goals and intentions (Csibra 2007, Hickok 2008 and 2014).

and sometimes even physically *enact* the urge to reorient one's own body to correct for the other's lack of balance.[2]

Notably, we do not only mirror actions. We also appear to mirror affectively laden sensations (e.g. pain and pleasure) and affective states such as emotions and moods. Upon seeing someone sustain an injury, for instance, human beings tend to represent both the intensity of the other's pain and the bodily location of that damage in their own sensorimotor systems (de Vignemont & Jacob, 2012). And, importantly, we appear to mirror not only sensorimotor components of pain, but the affective components as well. That is to say, when we mirror someone's painful experience, we do not just represent where it would be in our bodies, but how bad-seeming or unpleasant that feeling would be.[3] This may explain why we reflexively tense our muscles and prepare to pull away when seeing e.g. a syringe pierce another's arm.

There is also strong evidence that mirror neurons are responsive to emotional states via "emotional mimicry" or "emotional contagion." Quite often, the "contagious crying" is held up as a paradigm instance of this phenomena, wherein infants cry in response to the crying of other infants (Geangu et al., 2010). Remarkably, newborns cry significantly more in response to infants in distress than they do to white noise, synthetic crying sounds, or even recordings of the their own

[2] Importantly, all of this mirroring can happen swiftly and unwittingly. It does not require that an agent represent another's mind *as such*.

[3] In section IV, we will take a much closer look at what affect actually entails. For now, characterizing it as the "good/bad-seeming" or "pleasant/unpleasant" quality of any given experience will suffice.

crying (Sagi & Hoffman, 1976; Martin & Clark (1987).[4] This is not unique to newborns. Even adults spend their days subtly and inadvertently imitating the emotional facial expressions and body language of those around them (Chartrand & Bargh, 1999; Hess et al, 2011). Often, these bouts of mimicry seem to be accompanied by an *experience of* the emotion that is being mimicked, possibly through a mechanism of afferent feedback (Hatfield, 1994). On this account, when one mimics a friends' e.g. facial expressions of sadness, that physical mirroring results in the *feeling of* that friend's sadness or something close to it. Using fMRI data, Hennenlotter et al. (2009) seems to confirm this theory by demonstrating that amygdala activation (associated with emotion) is attenuated during imitation of others' facial expressions when Botox-induced paralysis prevent people from frowning.

In light of this research, mirroring and affective contagion certainly do seem like a basic kind of empathy, i.e. affective empathy. And it is not at all uncommon to describe them as such. In their own work on empathy, theorists such as Jesse Prinz and Paul Bloom explicitly define empathy merely as one's "ability to feel as others feel." As such, they would *ipso facto* include conscious action mirroring and contagion as instances of so-called empathy (Prisnz, 2011; Bloom, 2016). Similarly, Stueber refers to mirror neurons as mechanisms of "basic empathy," insofar as they seem to allow us to directly apprehend and experience others' mental states (Stueber, 2006).

[4] It is unclear, however, whether such cases are genuine instances of affective mirroring. When an infant hears the sounds of another agent in distress, that may very well be unpleasant and highly distressing. As such, the infants crying may be a result of personal distress and displeasure about the pain of others, i.e. it might be prosocial *concern* and not outright mirroring.

These processes do not, however, suffice for our more robust conception of empathy as *vicarious sharing of affect*—a characterization that is also commonly embraced throughout the literature.[5] Since the mirroring of basic actions has been observed in primates, songbirds (Prather & Mooney, 2014), and possibly canines (Range, Viranyi, & Huber, 2007), it presumably lacks sensitivity to other agent's context-setting e.g., beliefs, desires, explicit reasons, etc. That is to say, it does not demand or produce understanding of other agents' perspectives as such. This is also the case with "emotional mimicry." When an individual "catches" another's expression and feeling of e.g. joy, that contagion need not occur *in light of* the other's psychological states. In fact, pure "emotional contagion" does not even require that one consciously differentiate between self and other, nor *attribute* that emotion *to* another agent—one must simply *experience* the affect first-personally. As such, mirroring fails to suffice for a more robust conception of empathy, i.e. vicarious empathy.

B. PERSPECTIVE-TAKING

In order for one to vicariously share another's emotional state, it looks as though one must be able to not only feel as they feel, but *understand or represent that* they have tokened a particular affective state for a particular set of reasons. This is why work on mindreading is so crucial to the project at hand. In the theory of mind literature, "perspective-taking," "mentalizing," and "mindreading" all refer to one's ability to attribute mental states to others (Friedman & Leslie, 2004). This capacity

[5] Indeed, the *vicarious* affective sharing is especially relevant to moral philosophers, clinical psychologists, and cognitive psychologists alike.

is often described in everyday language as the ability to "put oneself in another's shoes." But, it is critical to note from the start that this characterization can be somewhat misleading. For, while we often think of "perspective-taking" as a deliberative or willful action—a *conscious effort* to "step into another's shoes"—imputing others' mental states is not always a conscious act.

Take the classic Sally-Anne task, a psychological test intended to measure one's ability to attribute false-beliefs to others. In paradigm iterations of this test, the subject is shown a skit involving two agents. Agent 1, Sally, takes a marble, hides it in her basket, and then exits the room. While she is away, Agent 2, Anne takes the marble out of Sally's basket and puts it in her own box. Sally then returns to the room, and the subject is asked "Where will Sally look for her marble?" If the subject has successfully represented Sally's mind, she should say that Sally will look where she last left the marble (and not where the marble is in fact hidden). In this task, subjects are tested on their ability to represent mental states such as goals, epistemic access or knowledge, and consequent false beliefs. (Baron-Cohen, Leslie & Frith, 1985). And, relevantly, one could easily expand this paradigm to include the attribution of affective mental states. If asked the question "How will Sally feel when she looks in her basket?" for instance, the subject should now be in a position to predict that Sally will be *surprised* or *dismayed* to find her marble is missing.

Although we are clearly capable of ascribing mental states to Sally in this scenario in a deliberative and conscious manner, our mentalizing is often driven by processes outside of conscious awareness. Indeed, the average person seems to impute Sally's mental states both unwittingly and without much effort or explicit

thought. This is why there is considerable consensus that mindreading can occur either consciously, slowly, and reflectively or unconsciously, quickly, and automatically, depending on one's circumstances (Sperber, D. & Wilson, D., 2002; Gallagher & Frith, 2003; Friedman, O. & Leslie, A.M., 2004; Kovács et al., 2010; Carruthers, 2008). How we should conceive of mindreading on the whole, however, and how we should characterize its outputs to conscious and unconscious awareness, is a matter of considerable controversy (for a review see e.g., Davies and Stone 1995).

There are, roughly, two large families of theories that attempt to model the causal mechanisms which allow us to interpret, explain, predict, and otherwise *understand* other agents. These are theory-theory and simulation theory. The theory-theorists posits that mindreading relies on the deployment of a "*theory"* or set of inferences about how mental states and agents operate; depending on the theorist, the mechanisms of that "theory" formation might be domain specific or domain general, and the theory might learned or innate, implicit or explicit.[6] Notably, theory-theorists are quite clear about what this process of mindreading yields us. On these accounts, mentalizing provides us with *information* about others' mental states. By way of illustration, imagine that while at the library, you see a student looking over his graded essay; as you draw near, a cursory glance at his papers reveals a large "F" scrawled in red ink, followed by a bevy of comments from the instructor. In this case, simply registering e.g. the student's failing grade and the comments in red, in tandem with knowing how people generally respond

[6] Since theory-theory is not the focus of this paper, we will leave questions of e.g. empiricism and nativism, modularity and domain-generality to one side.

to bad news, leads us to *infer* or *judge* that he is sad. Of course, we may (and often do) also neurologically mirror the student's sorrowful expression, a response that would presumably *feed into* this inferential process, but visual stimuli and mirroring are not necessary for mindreading on the theory-theorists account.

By contrast, simulation theorists argue that mindreading is achieved by cognitively "reenacting" the mental processes of others first-personally. On this view, one comes to represent others' mental states by simulating their circumstances, and then computing over those pretend or imaginary states *offline* as one would first-personally.[7] According to simulation theorists like Goldman (2006), mind-readers must then access the offline outputs of these simulations so that they can be projected onto the individual being simulated. To put it more systematically, in order to mentalize about our sorrowful student, one would need to (i) represent the particulars of his situation (e.g. seeing the bad grade in front of him), (ii) feed those representations *offline* into the appropriate cognitive mechanisms (those, for example, that generate emotional responses), (iii) directly access[8] the output of those mechanisms either consciously or consciously (representations of sorrow), and then (iv) attribute that output to the student.

[7] If simulations of e.g. affect were not usually offline, the story goes, they would presumably result in an immediate, first-person experience of emotion in that moment. In short, all simulationist mindreading would result in first-hand experience rather than representation of others' emotions.

[8] Not all simulation theorists think that we need to introspect these states per se. On a more radical version of simulation theory, Gordon (1995: 54) has argued that simulation is not "a transfer but a transformation." That is to say, we do not pretend to be the person we are mindreading, but rather indexically *become* that person. Whether simulation requires introspection or not, both accounts emphasize that we use our own cognitive processes to simulate other's minds, and that this is what lets us represent and understand them.

At this juncture, two things are worth noting. First—whether mindreading entails inferences, simulations, or both—when these processes are unconscious, they do not automatically give rise to the *experience of other's emotions*. In the case of theory-theory, the set of inferences that we use to deduce others' mental states need not result in anything like feeling what they feel. While tokened representations of the student as sorrowful may of course cause subsequent sadness, whether or not the mere representation will lead to *vicarious sharing* is an open question. The same is true of simulation theory. As stipulated by their own models, simulation requires one to register *that* one's own affective system is activated, often without feeling that activation as an occurrent emotion. This is likely one of the reasons that simulationists have placed such emphasis on mirroring (Gallese & Goldman 1998). In fact, some have characterized mirror neurons as the primary "automatic mechanism" by which simulation occurs, since they often seem to operate "below the threshold of consciousness, and can be detected only by brain-oriented techniques, commonly, functional magnetic resonance imaging (fMRI)." (Shanton & Goldman, 2010).[9]

Second, (and as a consequence of the first) it is apparent that mindreading must be engaged with *consciously* in order to give rise to vicariously shared affect. For, it is not enough to merely represent another's experiences (via an unconscious process), and then feel them for *whatever* reason.[10] Rather, vicarious sharing

[9] Though, as I have explained above and will reiterate later, I think there is good reason to suppose that mirroring is in fact often accessible to conscious awareness and can cause at least *shared affect* when triggered.

[10] One could, for instance, mentalize about a villain's failed attempt to take over the world and represent that he is sad, and then feel sad for unrelated reasons. Surely, this is not an act of empathy on any reading.

demands that we feel others' e.g. affective states in light of consciously accessible information about the generation of those feelings. That is to say, it is not enough to *know that* the student is sad, we must explicitly know *why* the student is sad and *what it would be like* (again, in conscious awareness) to feel as the student feels. Conscious access to these causal stories and phenomenology are clearly dependent on our mindreading capacities, but it is not only a mindreading process. Presumably, to meet the criteria set by vicariousness, one must actively mindread (again, whether via theory, simulation, or both) through acts of imagination and pretense.

A great deal of philosophical and empirical literature has explored the ways in which acts of imagination and pretense seem to engage the affective system and thus produce occurrent emotion (see e.g. Nichols (2004, 2006); Weinberg, & Meskin (2006); Meskin & Weinberg (2003); Schroeder & Matheson, 2006).[11] This is, in fact, the one area in which theorists have quite thoroughly explored the role of the affective system as it relates to empathy. Importantly, however, the affective system does not—as current models of empathy seem to suggest—simply *activate in response* to imagined input from our mirroring and mindreading systems. Rather, the affective system plays a crucial role in both triggering those systems and sustaining them. As we will see throughout the rest of this chapter, without affectively valuing others as worthy of concern or interest, there simply is no empathy.

[11] [11] Indeed, the relationship between imagination and emotion has generated quite a bit of controversy in Aesthetics. See chapter 4 for more on the "Paradox of Fiction." See also the "Puzzle of Imaginative Resistance" in e.g. (Moran, 1994); Gendler, 2000)

III. AFFECTIVE-APPRAISAL & UNDERLYING VALUES

The ability to psychologically distinguish between e.g., positive and negative, good and bad, pleasurable and unpleasurable, is fundamental to just about everything that we do. Valuing things in these ways impacts the decisions we make, what we pay attention to, the products we buy, who we fall in love with, what we find beautiful and ugly, etc. And, as we will see in the next section, whether or not we empathize with others (either through so-called affective or cognitive empathy) relies very much on how we value others and the world around us. But, what are the cognitive mechanisms that make these assessments possible? Where does this subjective valuing come from?

A. APPRAISAL AND VALENCE

In the previous chapter, we saw that the term "affect" includes a number of seemingly disparate psychological phenomena. These include emotions of joy, fear, and disgust; moods such as ennui, depression, and contentedness; sensory pleasures and pains such as those that accompany massages and toothaches; and desirous feelings like yearning and repulsion. We also saw that while there is some disagreement about what criteria might individuate these states, they share at least one thing in common: valence.[12] Valence is what contributes the phenomenal quality of goodness or badness, pleasantness or unpleasantness, attractiveness or aversiveness to all of these states (see Frijda, 1986; Ortony et al., 1990; Solomon

[12] They all also presumably involve concerted changes in (i) physiology, (ii) phenomenology (ii) and motivation. For a more thorough review of affect, see chapter 1.

and Stone, 2002; Russell, 2003; Charland, 2005; Colombetti, 2005; Barrett, 2006). So understood, valence is the mechanism through which our brains represent the positive or negative value of things in the world.

At a basic level of description, our ability to derive this value is driven by affective appraisal. Simply put, affective appraisal is the process through which one determines the subjective significance of any given situation, event, or object. Take, for instance, the experience of spotting a wild rhino while roaming around Serengeti National Park as the sun rises. If one is ill-advisedly roaming alone, and stumbles upon the large animal abruptly, one will likely appraise the situation as *dangerous*. Once that significance is computed, subsequent changes to physiology, motivational tendencies, and phenomenology (including negative valence, or bad-seemingness) will ensue, all of which contribute to one's overall feelings of fear or panic. If, on the other hand, one comes across the same rhino while safely tucked away in the armored vehicle of a guided wildlife expedition, one would likely appraise that situation as *a good opportunity*. After all, seeing a rhino in the wild is now an extremely rare and privileged experience! This appraisal would issue forth in different changes to one's physiology, motivation, and phenomenology (including positive valence, or good-seemingness), most likely to those states we associate with joy or excitement.

It is worth pointing out that in these scenarios, the appraisal system determined significance and produced valence along two different dimensions. In the first instance, the situation was appraised in terms of harm and benefit, i.e. the rhino was appraised as *harmful or dangerous,* which gave rise to *negative valence*

(the situation *seemed bad*). In the second scenario, however, the situation was significant because it was conducive to the appraiser's goals, i.e. the rhino was appraised as *an opportunity*, which gave rise to *positive valence* (the situation *seemed good*). Since the notion of affective appraisal was first suggested by Magda Arnold in 1960, many different structural dimensions of assessment have been proposed. Arnold herself argued that eliciting circumstances can be evaluated as good or bad, present or absent, and easy to attain or avoid. But, over the years these lists have grown. Lazarus (1991), for instance, proposed six different structural dimensions of assessment, including goal-relevance, goal-congruence or incongruence, type of ego-involvement, blame or credit, coping potential, and future expectancy. And Scherer et al. (2001) have gone so far as to propose that our appraisals arise along *sixteen* different dimensions, including e.g. moral significance.

Regardless of how one classifies and individuates these dimensions, it is clear that valence is produced in light of *some* standing considerations, including but not limited to things like safety, pleasure, social standing, etc. In the vast majority of cases, more than one of these considerations will come into play. When one walks out of the bush and stumble across the rhino standing meters away, for instance, one might appraise its horn as *dangerous* and (if the rhino is particularly unkempt) as *disgusting*. These appraisals bring with them different action tendencies (i.e. *run away* and *do not touch that,* respectively), but both appraisals activate the same regions of the brain associated with negative valence and thus *feel bad* (Levy & Glimcher, 2012). And, of course, one can also appraise a single situation along

different dimensions, even when the valenced outputs of those appraisals conflict wildly. From the safety of one's guided tour, one might appraise the rhino as *a good opportunity*, but also evaluate the situation as *morally bad* insofar as the rhino's solitary presence is a reminder of the species' decline.[13]

Certainly, all of the appraisals that we have mentioned so far *can* arise through conscious processing. One might consciously contemplate e.g., how seeing the rhino will fulfil a life-long dream, or how one's friends will be envious of these amazing adventures! But, appraisals can also be unconscious. Our evaluations do not just assign value to whatever we happen to be attending to: they also help *direct* our attention to things that are potentially relevant to us. In order for that to happen, the appraisal system must be sensitive to environmental cues that lie outside of focal awareness; these appraisals act as an "interruption mechanism" that alerts us to possible threats, sources of pleasure, opportunities, and so on (Kirby & Smith). Notice that as before, our conscious and unconscious appraisals may at times conflict. As one views the rhinoceros in relative safety, they may consciously appraise how *great* everything is, while nonetheless feeling a sense of subtle nervousness in the rhino's presence. In this scenario, one might consciously appraise that the rhino is *safe* (good), while unconsciously appraising the approach of a large, powerful animal to be *unsafe* (bad).

[13] To complicate matters further, it is also the case that for any single situation or object, the *targets* of one's appraisals may nonetheless be unique. One could, for example, appraise the rhino's form or shape as good on dimensions of attractiveness (its shape is beautiful), but its grimy horn as bad on the same dimensions (its horn is ugly). In one sense, the targets of these appraisals are different, but they can contribute to one's overall appraisal of the rhino.

The fact that we seem to hold different appraisals and clashing valence representations *at the same* time raises something of a puzzle for theorists working on affect. Namely, how is conflicting valence possible? As I explained in chapter 1, many researchers have characterized valence as a "common currency" that signals positivity or negativity. As such, valence is taken to represent a *single* axis of goodness and badness, one that gets instantiated across all instances of affect. And, as it stands, there is strong evidence in favor of this view. Empirical investigations into the matter reveal that all pleasant experiences, e.g. eating strawberries, listening to symphonies, earning extra money, being socially praised, etc., activate the same positive valence networks of the brain (for a review, see Cabanac, 1992). It has also been argued that valence must be a singular measure of goodness and badness, insofar as it enables us to make both conscious and unconscious decisions between options that seem otherwise incommensurate (e.g. deciding between staying home to eat strawberries or going to the symphony) (Levy & Gilmcher, 2012). But, returning to our working example, if the rhino-qua-threat is bad, the rhino-qua-opportunity is good, and the rhino-qua-moral concern is bad, how can we physically token all of that valence at once?

Some have tried to resolve this issue by claiming that we actually have multiple *kinds* of valence: one for e.g., money, another for power, another for pleasure and pain, etc., each corresponding roughly to different dimensions of appraisal (for a review of this debate, see Shuman, Sander & Scherer, 2013). While there is some evidence that different dimensions of appraisal arise along different neural pathways, this is no reason, I propose, to think of valence as occurring along

multiple axes. To begin with, positing different kinds of valence does not actually solve the problem of mixed emotions. After all, one might appraise the purchase of a new gold watch as monetarily bad-seeming in the short term, but monetarily good-seeming in the long run when viewed as a financial investment. Because these "money-valenced" appraisals would themselves conflict, we are still left with the problem of explaining how conflicting valence could co-occur.

My suggestion is this: valence *is* a single-axis "common currency," but it does *not*, in fact, co-occur. Rather, as our attention rapidly vacillates, the appraisal system outputs representations of valence successively—e.g. this rhino is *dangerous* (bad), this rhino is *novel* (good), this rhino is a *morally sad thing* (bad), then back to this rhino is *dangerous* (bad), and so on. Crucially, however, even if these valences are not simultaneously tokened, we should expect them to produce a fluid, overall sensation of "mixed feelings" in conscious awareness.

To start, when different dimensions of appraisal are engaged, we know that they result in differential activation of physiological and motivational tendencies. Appraising the rhino as *dangerous* (bad) will activate systems that prime us to flee, while appraising the rhino as *novel* (good) will prime us to approach. Importantly, though, each of these priming trajectories will remain *residually activated*, even as appraisals shift. Consider, for instance, what happens when you are startled by your beloved pet, Fluffy. At first, Fluffy's sudden appearance is appraised as unexpected and potentially dangerous (bad), thus triggering heightened physiological arousal and action tendencies to flee. Of course, you know that Fluffy is not dangerous, and shift your appraisals quickly…but, flight-or-flight tendencies can be quite slow to

extinguish. Even when you no longer actively represent Fluffy as *dangerous* (bad) in the moment, it can take a while for one's heart rate to slow and a sense of calm to return. These extended, bodily ramifications of appraisal will contribute to one's overall sense of an emotion persisting, even when valence has changed or flipped as the output of another appraisal.

Furthermore, although representations of valence might not co-occur, one's resulting experiences of good- and bad-seemingness can in principle be seamlessly integrated into one's overall phenomenology. By analogy, think of what it feels like to observe a fabric that shimmers with a green-blue hue. When we see this fabric, our eyes pick up wavelengths of light that typically correspond to one of two unique perceptual experiences, blue and green respectively. At any given point on the fabric at any given time, only one wavelength of light can be reflected; that reflection (and hence input to the visual system) changes continuously and in succession. And yet, despite this constant shifting of perceptual information, we are left with the *sensation* of looking at fabric that is *both* green and blue. This is, I propose, similar to what occurs in the case of valence. Because the appraisal process entails rapid-fire shifts in attention along multiple dimensions of appraisal, it likely results in rapid-fire vacillations of the valence network.[14] But vacillations in valence need not result in millisecond-by-millisecond changes in phenomenology: we can expect them to give rise to an overall feeling of mixed emotion that ebbs and flows over time. Moreover, our experiences of conscious valence (as felt through e.g. emotions) can be stored in working memory, such that even when they

[14] Unfortunately, given current technological limitations, we do not have clear enough temporal resolution in our measurements of affect to test this hypothesis empirically.

are no longer conscious, they can still impact our higher-order judgments about our own emotional states.

B. UNDERLYING VALUES

As we have just seen, dimensions of appraisal—e.g. safety, social standing, pleasure, morality, beauty, to name a few possibilities—set the bounds for which environmental stimuli will be appraised, and how they will be valenced. If the appraisal system were not sensitive to cues of e.g. safety and danger, then a rhinoceros' sharp horns would not elicit negative valence nor eventually result in feelings of fear. Described in this way, dimensions of appraisal seem to function as our *underlying values*. They determine both what grabs our conscious/unconscious attention, and how we will emotionally respond—whether we consciously articulate them or endorse them as worthwhile or not.

This means, importantly, that *underlying values* are not synonymous with "values" as the term is used in everyday speech. In ethics, for instance, the word "value" is often used to refer to one's preferences concerning appropriate courses of actions or outcomes, i.e. what *ought* to be. As such, a value can be thought of as a kind of abstract commitment. When one values e.g. friendship, beauty, or hard work, they have set an intention to pursue or embody those abstract concepts wherever possible (or at very least, to not overtly flout them).[15] But, of course, what we outwardly think ought to matter is not always what our brains (complex computational systems that they are) prioritize. A soldier might, for instance, find

[15] If someone is lacking in their commitment to e.g. hard work and thus rarely works hard, then one could reasonably object that they never "valued" it at all.

that an unwitting appraisal of danger and negative valence makes her feel terrified on the battle field, despite consciously endorsing or valuing (i.e. being committed to) courage. In this case, appraising the environment for signs of danger is an underlying value that might clash with the soldiers' relfective values.

This does not mean that our endorsed values have no impact on appraisal. Far from it! For, remember that for most of us, goal-congruence appears to be a highly sensitive dimension of appraisal, whether the goal is big or small, standing or occurrent. When one sets genuine intentions to clean the house, for instance, being thwarted in that endeavor *feels bad* (even when one is thwarted by pleasant circumstances, like an unexpected visit from a friend). If values are kinds of commitments, then breaking with those commitments should also trigger an appraisal. In this case, it would generate an appraisal of *non-goal-congruence* (bad) and, if bad enough, would motivate one to change their behaviors. Furthermore, some theorists have argued that we naturally appraise the world along dimensions of self-(in)congruence and moral badness/goodness. If this is true, then not acting in accordance with one's consciously endorsed values would certainly suffice to trigger an affective appraisal, likely one resulting in e.g. shame or disappointment.

Whatever the dimension of appraisal, this emotional akrasia (feeling bad about our feelings) is often potent enough to make us change course. In the literature on willpower, several empirical investigations have revealed techniques that enable us to successfully regulate our emotional appraisals (Metcalf & Mischel, 1999). If our soldier is compelled to subdue her fear on the battlefield, for instance, she might (i) distract herself from whatever is generating an appraisal of threat, or

(ii) reframe it. To (i) distract herself, the soldier might attend to the grass, thoughts of chocolate, memories from her childhood, i.e. she could think of anything other than what is causing her fear. To (ii) reframe things, she might imagine, for instance, how her peers will admire her bravery, or how sad and disappointed her family will be if she lets them down: she is still attending to the initial target and cause of her fear (impending battle), but is now deliberately trying to do so in a different light.

Why do these techniques work? In both cases, it looks as though providing the appraisal system with new stimuli (thoughts of chocolate or social approbation) activates *different* dimensions of appraisal. Chocolate is good-seeming on dimensions of pleasure, and social approbation is good-seeming on dimensions of social standing. Affect is still being tokened, it has just shifted in dimension and valence. By consciously changing what we attend to and how we attend to it, we are often able to shift dimensions/modes of appraisal, and this brings our emotions in line with our more abstract values. In this sense, our underlying values and higher order values are deeply interconnected.

IV. EXPLAINING FAILURES OF EMPATHY

It should go without saying that the affective system is necessary for empathy. It is, after all, the very system by which we *feel* emotions. In much of the contemporary literature on empathy, however, affect is either treated as an obvious add-on to be explored in future expositions, or as an automatic consequence of engaging with others' minds. Many seem to imply that we mirror others and *then* feel what they

feel; that we imagine what it is like to be someone or pretend to be in their shoes and *then* the affective system is activated. This may very well be true…but it is not the entire story. There are many instances in which people (both those who have been clinically diagnosed with antisocial tendencies and those who have not) can mirror and mindread, yet fail to empathize with others. At present, most of our theoretical models of empathy simply do not account for these failures. Now that we have taken a look at how affective appraisal works and the nature of valence, however, we stand in a much better position to explain why this happens.

A. Psychopathy

It has long been acknowledged that psychopaths lack empathic tendencies. Indeed, one of the strongest criteria of diagnosis for psychopathy in clinically-endorsed psychological assessment is a general callousness towards others' feelings and an absence of prosocial sentiment, be it empathy or sympathy (Venables et al, 2013). But what, exactly, is happening at the level of cognitive architecture to cause this lack of empathy? In 2008, Shirley Fecteau et al. found that non-psychiatric psychopaths[16] are quite capable of mirroring the location and intensity of other's pain. Surprisingly, in fact, those who showed the greatest propensity for "cold-heartedness" or lack of empathy showed *greater* sensorimotor mirroring than did the average participant. What these subjects did not do, however, was to "catch" or feel the affective component of the pain that they observed.

[16] If these were the results for non-psychiatric psychopaths, it is safe to assume that those meeting criteria for diagnosis would exhibit these tendencies even more robustly.

Why do psychopaths mirror the *sensation* but not the *valence* of others' pain?[17] At present, most theorists characterize affective mirroring as an "automatic mechanism" that leads from: (i) perception of affective cues (a frowning face, a needle's injection, etc.); to the (ii) mirroring of those cues (activation of neurological regions associated with frowning, the needle's injection, etc.); to (iii) the tokening of action tendencies (flinching) and felt valence (presumably negative). [18] Clearly, that process is not automatic in the case of psychopaths. Nor, I would argue, is it automatic for anyone. In order for mirroring of sensorimotor damage to give rise to valence, it has to be *appraised* as bad. For most of us, in our everyday dealings, mirroring other's bodily harm simultaneously triggers an appraisal of badness. This is because most of us really do *care* about others and see their suffering as bad. But, we don't just value or care about others in abstraction: for the average person, others' wellbeing is a critical underlying value to which the appraisals process is highly sensitive. Conversely, when the psychopath mirrors sensation and not valence, it seems to be because she lacks this underlying value and as such does not appraise others' pain to be bad. This breakdown in affective empathy is thus not the result of malfunction mirroring, it is the product of one's appraisal.

[17] As we discussed in the previous chapter, affective states seem to have both affective (valenced) and bodily (sensorimotor) components. Morphine, for instance, has been shown to drastically lessen the seeming-badness of physical damage to one's body while leaving intact one's representation of damage in the sensorimotor cortex. For more on this, see chapter 1.

[18] One might object that mirroring emotion via a cue like the prick of a needle is too abstract, i.e. that it requires the psychopath to first recognize the other's pain and then feel it. This seems unlikely, since psychopaths actually perform slightly better than controls on some complex emotional recognition tasks. (Dolan and Fullam, 2004).

Psychopaths also have been shown to perform on par with normally functioning adults on mindreading tasks (Dolan & Fullam, 2004). Further, Decety et al. (2013) have demonstrated that they appear perfectly capable of imaginatively taking the perspective of others to deduce their emotional states; they simply do so in a manner that is remarkably un-affective. This research shows that when imagining their *own* future states, psychopaths demonstrated considerably more activation of the anterior insula and amygdala with the orbitofrontal cortex and ventromedial prefrontal cortex—regions of the brain associated with valenced appraisal and affect—than when imagining others' experiences. That is to say, while psychopaths are able to represent and attribute others mental states, they do so without vicariously sharing in their affect.

Importantly, this is not a failure of the appraisal system per se. As is made clear by this study, psychopaths do not simply lack appropriate affective appraisal across the board. Indeed, individuals diagnosed with psychopathy often display narcissistic valence and arousal patterns; they care deeply about anything that is in their own self-interest, e.g. they experience intense anger in response to frustration, feel exaggerated personal distress, etc. (Blair, 2012; Martens 2002). When imaging others' experiences, however, the psychopath does not appraise those experiences as good or bad, i.e. they are affectively flat. Two things might account for this flatness. First, it is possible that because the psychopath does not value others (in either the traditional or underlying sense), they do not bother to furnish their imaginative simulations with the kinds of rich details that trigger the affective system (see e.g. Moran, 1994). Or, second, it is possible that psychopathic subjects

did represent the other's experience in a sensorially rich format, but that those representations were simultaneously tagged as not pertaining to them, and hence did not trigger the subject's underlying values for affective appraisal.[19]

B. Sadism

Sadism is also associated with a lack of empathy. While it has been removed from more recent editions of the DSM, sadism is often described as a personality disorder in which one takes pleasure from the negative affective states of others, e.g. discomfort, humiliation, pain, etc. Considerably fewer empirical studies have been conducted on sadists' capacities for mindreading, mirroring and the generation of affect. That said, there are compelling reasons to believe that sadistic individuals understand others' emotional states via the same cognitive mechanisms as non-sadistic individuals. Indeed, it seems plausible that the sadist's preference for suffering in others *requires* the ability to richly calculate, represent, and understand those emotional states; were sadists unable to do so, they would have nothing to get pleasure *from*. In one of the few clinical studies conducted on sadism, Harenski et al. found that when viewing pictures of painful experiences, sadists showed greater activation of the amygdala (a brain area associated with strong emotion) than did other sexual offenders. As compared to non-sadists, these sadists also showed greater correlation between the pain severity ratings of the photos and activity in the anterior insula: the more pain the sadist attributed to others in a particular

[19] If it is possible for the psychopath to imagine someone's experience in vivid detail and not experience affect, that is noteworthy and stands in need of explanation. If underlying values or dimensions of appraisal really are what make us emote, then it is possible that psychopath's underlying values concern only what is relevant to themselves.

scenario, the stronger the affective reaction. These findings suggest that (quite unlike the psychopath who is predominantly indifferent to the suffering of others) the sadist is highly attuned to others' experiences.

It seems likely, then, that the sadist has access to all of the same informational content (accrued through sensorimotor mirroring and perspective-taking) as anyone else. What distinguishes the sadist from non-sadistic individuals is how they appraise that information. As we have established, representations of pain in others usually triggers swift appraisals of badness. When sadists detect that another is in pain or otherwise subjugated, however, that input to the appraisal system results in positive valence. But, how is this possible? The underlying value or appraisal dimension that is sensitive to pain and pleasure *for oneself* presumably functions typically in people who systematically enjoy harming others, i.e. they do not flip their appraisals of pain and pleasure and become masochists themselves. So, why appraise others' pain as good?

I suspect that unlike psychopaths, sadists do not fundamentally lack an underlying value that is sensitive to other's wellbeing. Sexual sadist, for instance, derive pleasure from others' suffering, but presumably do so only in the context of sexual acts; they are still capable of empathizing with others in their daily lives. So, why do sexual sadists not feel the pain of others in sexualized scenarios? Some have proposed that human beings are naturally inclined to evaluate the world on dimensions of power, control, or personal agency (Scherer, 2009). That is to say, some claim that the appraisal system is constantly scanning for cues that indicate one's social rank, personal influence over others, access to resources, and so on. I

suggest that sadists actually appraise others' suffering in light of *this* underlying value. For whatever reason,[20] sadists appraise others' expressions of pain not primarily as cues for moral or social appraisal, but power. And, in some respects, the sadist is correct; causing pain in others *is* a signal of power over others, and power is generally appraised as good. For the vast majority of people, however, pain in others triggers appraisal along dimensions of moral and social badness, and so we do not appraise their pain as *power* (good)—quite the contrary!

C. Everyday Failures

As we have seen, theorizing about the nature of affective appraisal allows us to better explain those with systematically aberrant empathic tendencies. But, this does not just apply to atypical populations. At some point or another, all of us will fail to empathize or experience what is known as an "empathy gap." I contend that this happens, at least in part, because our own self-interests generate affective states that can permeate and counteract the affect we token when imaginatively simulating and mirroring others' experiences.

To see why this is the case, consider the fact that receiving an injection one considers to be useful is often reported to *feel less painful* than injections which serve no apparent purpose at all (Leknes, 2013). Why does this happen? When an injection is appraised as a source of occurrent damage, the sting of the needle will quickly produce *negative* valence. However, when an injection is appraised solely in light of its ameliorative effects, that potentially life-saving medical procedure

[20] See final section for an idea of why the sadist might differentially appraise cues like expressions of pain.

looks quite *positive*. Importantly, the appraisal system will iteratively appraise and reappraise the injection in light of (at least) these two dimensions as it is being administered, resulting in the generation of both positive and negative valence. Since these representations of valence are both projected onto the same target, i.e. receiving the injection, one's overall feeling of pain is mitigated.

I suspect that this mitigation of affect is rather common when we attempt to engage in cognitive empathy. Suppose my colleague has recently received a promotion, and I sit down to imagine what that experience was like for her. When I imagine this situation from my friend's perspective (knowing how hard she has worked, how much she wanted to be recognized), it evokes satisfaction and contentment. In this case, assuming my friend is as happy as I think she is, I have successfully empathized. But, much can impede this process. Suppose now that my company famously only gives out one promotion per year. While all of this is still good news for my colleague, it might greatly frustrate my own goals and sense of self-worth, making empathy difficult. After all, when one imagines or pretends to be another, one is all the while aware that they are engaging in an act of imagining or pretense, i.e. one knows that they are appraising things as good or bad *for another person*. As I imagine what it is like to shake my boss's hand through the perspective of my overjoyed colleague, I am still happy *as her*…but this new information makes intermittent appraisals of goal-incongruence and power loss much more likely.[21]

[21] This is, likely, why it can literally feel bad to try to empathize with those we find morally repugnant or villainous. While we can imagine what it is like to be a terrible despot living in the lap of luxury, understanding that that simulation is *true for someone evil* will generate its own affective appraisal.

In short, when we endeavor to empathize, we do not just appraise the contents of our imaginations. We also appraise those contents for what they represent at large and how they might impact us as individuals. If the valence of those disparate appraisals contradict one another (as was the case with the injection), it will undoubtedly shift the emotions we experience, making shared affect unlikely. All of this is to say that whether we take on the *same* appraisal of another's circumstances as they do depends in heavy measure on the relationship that they bear to us and how their circumstances impact our own underlying values, e.g. our own safety, goals, pleasure, and so on.

D. Intra and Intergroup failures

Our underlying values around group membership seem to have an especially strong impact on mirroring and perspective-taking tendencies. In 2010, Avenanti et al. found that group membership along racial lines had a profound impact on the mirroring of others' pain. In this study, researchers showed participants videos of a "neutral" hand that was colored purple, the hand of a black person, and the hand of a white person, each of which was being pricked by a needle. Remarkably, participants with strong implicit racial biases mirrored the sensorimotor pain of the purple hand, and the hand corresponding to their own race, but not the hand for which they held implicit bias. Similar effects have been found in cases of "emotion contagion," where i.e. mirroring expressions of fear, happiness, anger etc. appears to be mitigated by the extent to which one identifies with another's group (Hess et al, 2011). While it is rather unfortunate that we have

these tendencies, it looks like actively appraising others as bad-seeming along dimensions of, say, social standing or social membership, is enough to prevent what otherwise seems to be an automatic mirroring process.

Over the past few decades, a growing body of research has shown just how impactful these appraisals of group status are. Within their first year of life, infants tend to show preferences for imitating those who speak the same language and exhibit the same appearance (usually along racial lines) as their caregivers (Buttelmann et al., 2013:). By age five, these preferences turn into a tendency to both allocate more resources to in-group members, and to form more positive social expectations of those members, i.e. those who belong to one's in-group are assumed to be more cooperative and better behaved (Dunham et al., 2011). Remarkably, this preferential appraisal of outgroup members can be impactful even when groups are novel, i.e. when membership is assigned randomly. And, especially germane for our purposes, self-reported data indicates that when presented with photographs of painful experiences, people are more inclined take the perspective of those who they identify as belonging to the same racial group, particularly when observing negative emotions. (Neumann, Boyle, Chan, 2013).

V. Conclusion & Future Research

At this point, we have seen that the affective appraisal system plays a fundamental role in the success and failure of empathy. By incorporating some of our best empirical and theoretical accounts of affect, I hope to have elaborated on our models of what it takes to feel as others feel, vicariously or otherwise. We have

seen, for instance, that having a dimension of appraisal that is sensitive to others' well-being is a critical component for empathy: it impacts when we mirror and take others' perspectives, and it sets our overall disposition to care for those around us. But, for all that this emphasis on affect illuminates, it also raises a number of questions. The most pressing of these questions revolves around how dimensions of appraisal are set and weighted. After all, if underlying values do all that they seem to do, then knowing how they are formed and interact is of the utmost importance.

While I do not purport to answer these questions here, one area of research that might shed light on these questions is that which focuses on "primary" and "secondary" reinforcement. Primary enforcers are objects or conditions for which an animal has strong, genetically determined associations of pleasure (understood here as positive valence). These primary reinforcers are thought to be innate, and usually fulfil some biologically necessary function, such as food, shelter, sex, etc. By contrast, secondary reinforcers acquire their value through a history of learned association with already-valued primary reinforcers or other valued secondary reinforcers. To use a classic example from the literature, when one gives a dog a treat and tells him that he is a "good boy," the dog gets both a *primary* reinforcer (the treat, which innately tokens positive valence) and a *secondary* reinforcer (verbal praise, which is not innately valued by the dog as such). Once one's verbal praise is associated with the positively valenced experience of food, that verbal praise begins to token positive valence for the dog, even in the absence of treats.

As we saw in section III, appraisal theorists have proposed a number of appraisal dimensions or underlying values—many more than food, water, and sex. That said, most of the dimensions that have been proposed do seem, if not necessary for human survival, *optimal* for survival e.g. safety/threat, pleasure/displeasure, goal-conduciveness, power status, social/moral standing, etc. And, since human beings are undoubtedly affective little creatures from the moment of birth, it makes sense to assume that we have some innate underlying values that help us navigate from the start. Indeed, infants seem disposed from very early on to e.g. respond fearfully to snakes and spiders (Hoehl, Hellme, Hedqvist, & Gredebäck, 2017), find sweet things to be pleasant (Steiner et al., 2001), plainly get frustrated when their own goals or wants are stymied, and are attentive to both social hierarchies (Thomsen, 2019) and prosocial/antisocial behaviors (Hamlin, 2013). It seems reasonable to suppose, then, that these dimensions of appraisal are present at birth and are disposed to generate the valence of a number of set primary enforcers in predictable ways.

There are, admittedly, some limitations to describing human preferences in these terms. For instance, it is not entirely clear how *primary* our "primary reinforcers" really are. One the one hand, research suggests that evolutionary pressures may have hardwired the appraisal of particular stimuli to be good- or bad-seeming, especially when they directly impact our survival as a species. Sugar preferences, for instance, likely evolved because sugar is an easily detectable, roughly accurate signal that something is edible (Ramirez, 1990). As such, sweetness seems to be liked by just about everyone. On the other hand, once we are

older and more cognitively developed, the very same stimulus—even that which appears to work as a primary enforcer, like sugar—can be appraised on different dimensions. So, although sugar triggers appraisals of sensory pleasure and positive valence for *most people*, it could easily trigger appraisals of e.g. goal-(in)congruence if one were on a diet, or e.g., moral-wrongness if one were, say, suffering from an eating disorder that made the consumption of food feel morally repugnant. In these cases, those who do not get pleasure from sugar may still have an underlying value that is sensitive to pleasure *in general*, sugar now simply fails to be appraised along that dimension.[22]

These worries aside, further exploration of research on e.g. associative learning in tandem with what we know of the appraisal system is of crucial import. Hopefully, by setting our sights in this direction, we will get a better grasp on why, for instance, the sadist appraises others' pain as a signal of power, rather than a cue for social or moral distress. Or why, for instance, the psychopath seems to either lack or override underlying values that would otherwise lead her to care about others. Only when we gain traction with some of these unaddressed questions will we have a comprehensive theoretical model of how empathy works.

[22] The same seems to be true of the sadist and her association of other's pain with her own power.

CHAPTER 3

DISFLUENCY, NEGATIVE AFFECT, AND INTEREST: A BETTER THEORY OF AESTHETIC PLEASURE

I. INTRODUCTION

Human beings ascribe beauty to an immense variety of things: masterful dances, haunting melodies, inviting landscapes, charming faces—the list is virtually inexhaustible. And yet, despite the fact that judgments of beauty abound, they remain remarkably enigmatic. What, exactly, is a judgment of beauty? What drives our feelings of aesthetic preference and pleasure across such a broad and diverse spectrum of stimuli? Why do we deem some objects beautiful, while others altogether fail to inspire appreciation?

Of course, these are by no means new questions. Plato, Aristotle, and their western contemporaries famously began theorizing about aesthetics over two thousand years ago; since that time, we have produced an extensive body of literature exploring beauty's ontological, moral, and epistemic status. In recent years, however, the growing field of empirical aesthetics has provided us with a novel, more scientific vantage point from which to consider these long-standing quandaries. Unlike strictly philosophical approaches, the empirical aesthetics project is aimed at discovering the evolutionary and cognitive foundations of phenomena like aesthetic pleasure and preference.[1] To that end, researchers in

[1] Whether we can reduce judgments of *beauty* to feelings of aesthetic *pleasure* and *preferability*—the latter of which I take to be appraisals of "seeming goodness—is a subject of some debate and warrants its own treatment. For my purposes here, I will assume that subjective feelings of aesthetic pleasure and preference are likely intrinsic to the process of

empirical aesthetics often employ a three-fold, interdisciplinary approach. By considering relevant scientific findings in e.g. cognitive and evolutionary psychology, neuroscience, experimental aesthetics, etc., theorists intend to discover the conditions under which we are likely to make judgments of beauty (what we like), form a plausible theory as to why these preferences and judgments of beauty have evolved (why we like it), and deduce how those judgments are made at a cognitive level of description (how we like it). The hope is that by evaluating philosophical and scientific theories of beauty in tandem, we will be able to develop a more holistic, falsifiable, and psychologically explanatory model of our everyday aesthetic judgments.

Over the past several decades, a number of promising but incomplete theories of aesthetic pleasure have been put forward. Daniel Berlyne's highly influential "arousal potential" view, for instance, posits that our strongest aesthetic preferences are directly correlated with the extent to which a stimulus causes not low or high, but moderate psychological arousal (Berlyne, 1971; 1974). On this account, a mildly complex or arousing sensory experience generally indicates that an organism's environment is advantageous, and so should produce the strongest feelings of approbation. By contrast, Colin Martindale and Kathleen Moore challenge Berlyne's take with a more cognitive explanation of aesthetic pleasure. Their "prototype preference" theory holds that aesthetic pleasure is not a measure of complexity or arousal potential per se, but of a stimulus' typicality. That is to say, they posit that the more clearly and concretely a subject is able to conceptually

judging an object as beautiful (though, for a more skeptical account, see e.g. Armstrong & Detweiler-Bedell, 2008.)

classify a perceptual stimulus, the more pleasing she will find it (Martindale & Moore, 1988).

While these theories certainly elucidate *some* of our tendencies to regard stimuli as beautiful or pleasing, they unfortunately fall short of accounting for the aesthetic experience of beauty *as a whole*. To put the problem rather succinctly, each approach seems to fail explanatorily where the other one succeeds.[2] Recently, however, the "processing fluency" theory of aesthetic pleasure has been suggested as a means both to elaborate upon what these views get right and to unify them under a single framework. To date, Reber and his colleagues have provided the most comprehensive articulation of the processing fluency theory, which holds that the pleasantness of aesthetic stimuli can be identified with the perceptual fluency those stimuli afford (see Reber, Schwarz, & Winkielman, 2004; Reber, 2012). In this context, perceptual fluency refers to the subjective experience of *effortlessness* one has when engaging with the perceptual features of a stimulus e.g. its form, color, size, etc. According to proponents of this view, when an object's sensory properties are processed with ease, we tend to feel positive affect which drives our subjective ascriptions of beauty; when we struggle to process an object's sensory properties, we find that object less pleasing and hence less beautiful.

As I will show, Reber et al.'s work provides a model of aesthetic judgment that aligns better than most with a wide breadth of scientific findings on sensory preferences, subjective feelings of liking, and art appreciation. But, while their model fits well with much of the data, it is currently open to a couple of critical

[2] I will expand upon this claim further in the section that follows.

objections that need to be addressed. Namely, the fluency-as-pleasure view (i) fails to truly explain why people tend to dislike stimuli that afford extremely fluent processing—in Reber et al's attempts to do so, they rely on problematic *ad hoc description* rather than theoretical *explanation*—and (ii) it discounts the impressive flexibility with which both art experts and laypeople are able to shift their appreciative stances. These issues arise, I will argue, primarily due to the fact that Reber et al.'s account ignores the crucial role of affectively-laden *interest* in eliciting aesthetic pleasure.

The following is intended to amend this oversight. In sections II and III respectively, I will lay out what the processing fluency theory of aesthetic pleasure successfully accomplishes, and then highlight some of its most evident shortcomings. Once I have established both why this view is worth building upon and the challenges that it faces, in section IV I will focus on developing two new claims. First, given contemporary work on the affective mechanisms driving e.g. attention, curiosity, problem-solving, etc., I will propose that interest qua primitive "questioning attitude" is necessary for genuine aesthetic appreciation. Second, it seems likely from an empirical standpoint that this questioning attitude often supervenes upon processing *dis*fluency and resulting *negative* affective states which motivate, amplify, and sustain ongoing aesthetic pleasure. Together, these elaborations should give us a more comprehensive model of affectively-laden perceptual processing with the explanatory power necessary for rectifying the processing fluency view.

II. POSITIVE FEATURES OF THE ACCOUNT

Providing a robust explanation of aesthetic pleasure is no simple task. Any worthwhile, scientifically-based theory of beauty must accommodate not only a diverse set of experimental data, but also account for a number of well-vetted philosophical observations about the nature of aesthetic experience. Prior to Reber et al.'s processing fluency view, two views about why and how we form such judgments were especially influential. By drawing from different scientific fields of research, Berlyne's arousal-potential theory and Martindale and Moore's prototype-preference theory provide unique explanatory schemata for a wide range of findings. As I will argue, although both of these theories leave something to be desired, they nicely illustrate the spectrum of phenomena that any complete theory of aesthetic pleasure must explain.

The first of these, Berlyne's arousal-potential view, posits that our aesthetic preferences are correlated directly with a stimulus' capacity to cause arousal. In cognitive psychology, arousal is roughly characterized as the degree of physiological and—of particular import for our purposes here—*psychological* excitement that an organism experiences when confronted with a stimulus.[3] The arousal-potential view builds upon a well-established principle in the domain of evolutionary psychology; namely, that high and low states of arousal provide aversive motivational feedback to an organism. For instance, a preponderance of

[3] Compare, for instance, the phenomenal contrast between contentment and fear. When content, an animal's physiological and psychological states are relatively close to baseline, e.g. its heart rate is regular, breathing is steady, and the animal is able to shift attention with relative ease; by comparison, when in a fearful state, the same animal is likely to be agitated both physically and mentally, e.g. its heart rate climbs, breathing will quicken, and its attention homes in dramatically on any potential source of negative affect (see e.g. Easterbrook, 1959*)*.

highly arousing stimuli has a strong likelihood of occurring in turbulent or dangerous environments, while a dearth of arousing stimuli may indicate that an environment will provide suboptimal resources for survival. If this is the case, then an environment that is relatively safe yet fruitful and worth exploring should produce *moderate* psychological arousal and an accompanying sense of strong approbation. With this understanding in place, Berlyne reasonably concludes that our most intense feelings of aesthetic pleasure must have evolved to track middling arousal vis-a-vis moderate perceptual complexity and novelty (Berlyne, 1971; 1974).

The arousal-potential view both succeeds and fails on different dimensions. It aligns nicely, for instance, with the empirical finding that stimulus complexity is often related to preference by an inverted U-shaped function, i.e. it aligns with the average person's preference for sensory experiences that are neither too simple nor too complex, neither totally alien nor overly familiar, etc. (Berlyne, 1971; Vitz, 1966). Reducing aesthetic pleasantness to arousal-potential also allows us make sense of the fact that several objective sensory properties appear to be commonly favored in the literature on subjective liking. In a number of studies, it has been found that subjects prefer perceptual features such as visual clarity (Whittlesea, Jacoby, & Girard, 1990), symmetry (Reber, Schwarz, & Winkielman, 2004; Wurtz, Reber, & Zimmermann, 2008), and figure-ground contrast (Reber, Winkielman, & Schwarz, 1998; Thompson & Ince, 2013), all of which push stimuli into a mid-range of complexity and novelty. But, this view also has considerable problems. As Martindale noted, arousal potential alone cannot explain why two sources of stimuli

that are *equally arousing* can be *differentially appraised*. After all, if Berlyne's view fully captured our preference tendencies, then a discordant cacophony should be just as pleasurable as a spirited melody of the same arousal-potential, but that does not seem to bear out in everyday experience—*how* we engage with sensory input matters as well (Martindale, 1988).

Martindale & Moore suggest that a more subjective, "cognitive" process is involved in shaping our attributions of beauty. Their prototype-preference theory asserts that aesthetic pleasure is not driven by arousal per se, but by a stimulus' fit to central tendency of a category. In other words, the more archetypal the subject finds a stimulus, the more that stimulus is liked (Martindale & Moore, 1988).[4] Many studies have revealed this prototype-preference effect in aesthetic contexts. Research demonstrates, for instance, that perceptual properties such as color typicality can influence liking to a much greater extent than mere figure complexity (Martindale, 1990); subjects also appear to prefer paintings that are typical for a given artistic style, even when controlling for factors such as novelty and familiarity (Farkas, 2002). But, while the prototype-preference theory accommodates these findings and avoids some of the pitfalls of the arousal-potential model, on the whole its explanatory power is limited. We are still left to wonder, for instance, why one would prefer prototypes in the first place; is there an evolutionary or cognitive account of why this tendency might come about? Further, the prototype-preference view struggles to explain the U-shaped function between complexity and

[4] On this view, when evaluating dog breeds, the average Westerner might prefer Labradors over Bull Terriers because the former more clearly resemble the abstract average of all of the members of our culturally shared "dog" category.

preference. For instance, it is not hard to imagine a stimulus that is extremely simple because of its strong typicality, and hence disliked for that very reason—yet such distaste should not be possible if we merely favor prototypes. Finally, a theory that reduces aesthetic pleasure to prototypicality has a rough time explaining the appreciation we often feel for a staggeringly broad range of stimuli e.g. non-representational art, abstract patterns, works that challenge genre expectations, etc.[5]

Although the arousal-potential and prototype-preference theories of aesthetic pleasure are imperfect, they have both added considerably to the empirical project at hand. By highlighting the fact that both objective sensory qualities (e.g. complexity or arousal-potential) and subjective cognitive processes (e.g. the detection of prototypicality) contribute to our feelings of aesthetic pleasure, they pave the way for a more comprehensive, unified theory. Recently, Reber et al. have led the charge in providing that more comprehensive theory with their processing-fluency model.

Reber and his colleagues suggest that aesthetic pleasure is ultimately a function of the perceiver's processing dynamics (e.g. Reber, 2012; Reber & Schwarz, 1999; Reber, Schwarz, & Winkielman, 2004; Reber, Winkielman, & Schwarz, 1998). Specifically, they argue that a stimulus is experienced as pleasing or beautiful when it affords optimal levels of processing fluency. In this context,

[5] Indeed, as others have mentioned, one shortcoming of the prototype-preference theory is that "typicality" is hard to define beyond the particulars of a specific act of aesthetic appreciation (Silva, 2012). An artwork, for instance, could be typical or atypical with respect to its e.g. diverse formal properties, representational content, intended historical style, etc.—to simply state that aesthetic pleasure is a measure of "typicality" *simpliciter* is to ignore the fact that typicality can only be established within the context of cognitive framing (for a thorough philosophical treatment of this topic, see Walton, 1970). I will return to this topic in sections III's discussion of questioning attitudes.

processing fluency is best characterized in terms of the subjective feelings of *ease* or *difficulty* one has when processing any sort of information (Reber, Wurtz, & Zimmermann, 2004).[6] In their account of aesthetic appreciation, however, Reber et al. focus almost exclusively on *perceptual* processing fluency, or the level of effortlessness we experience when engaging with strictly perceptual qualities e.g. form, size, tone, pitch, or any other sensory detail (Graf, Mayer, & Landwehr, 2017). Importantly, Reber has argued elsewhere that fluent processing such as this is hedonically marked, i.e. fluent processing intrinsically feels good while disfluent processing feels bad (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Drawing upon the same scientific hypotheses as Berlyne, these theorists claim that fluent information processing results in positive affect because "high fluency signals…that things are familiar and the ongoing cognitive processes are running smoothly, whereas difficulty of ongoing processing signals that things are not going well." (Reber, 2012, p.227). Putting all of these observations together, the processing fluency view of aesthetic judgments holds that we find an object beautiful because the fluent processing thereof produces pleasure, which is then attributed to the object itself.

The processing-fluency theory does a very good job at grappling with phenomena that previous theories have only captured in part. Since complexity,

[6] Because it gives rise to a phenomenal state, fluency can be gauged by subjective measures, e.g. subjects' ratings of a task as easy or effortful; more often than not, however, it is gauged via objective measures of speed and accuracy, e.g. pronunciation latency (e.g., Whittlesea, 1993; Whittlesea, Jacoby, & Girard, 1990), naming latency (Reber et al., 1998), or likelihood of identification (Jacoby & Dallas, 1981). The assumption here is that objective measures of fluency generally coincide with subjective feelings of fluency. For more on this topic, see Reber, Wurtz, & Zimmermann (2004).

novelty, and prototypicality all have an impact on processing fluency, it is easy to subsume these theories under the framework that Reber et al. have provided. Like Berlyne's arousal-potential theory before it, the processing-fluency view nicely accommodates the fact that we often feel distaste for overly-complex stimuli: since extremely complex perceptual properties are by their nature difficult to process, Reber et al.'s model rightly predicts that they will not produce the feelings of fluency required for aesthetic pleasure. The processing fluency view also provides a more structural explanation of our preferences for e.g. familiarity, visual clarity, symmetry, and figure-ground contrast; these properties facilitate ease of processing, and so should predictably be preferred. The processing-fluency theory also goes a long way towards expanding upon the prototype-preference view insofar as it explains *why* prototypicality affects preference as it does. After all, it makes sense that central members of a category would be preferred by virtue of the fact that they are easier to process than members at the fringe (Winkielman, Halberstadt, Fazendeiro, & Catty, 2006). Finally, the processing-fluency model provides us with explanations for other well-documented findings in cognitive psychology and experimental aesthetics. For instance, it offers a particularly strong account of why priming has a positive impact on judgments of liking, since priming makes subsequent processing more fluent and thus pleasurable (Winkielman and Fazendeiro, 2001).

III. PROBLEMS WITH THE PROCESSING-FLUENCY ACCOUNT

The fact that processing-fluency succeeds in providing a unifying framework for such a diverse array of research and observations suggests that it is on the right track. Indeed, it is the hallmark of any strong scientific theory that it explains and is supported by as many independent strands of data as possible. But, while the processing-fluency view of beauty accomplishes a lot, it is currently vulnerable to several serious criticisms.

The first, and perhaps most often voiced, criticism against the processing-fluency view is that it struggles to robustly explain why people tend to dislike stimuli that are extremely simple. If fluent processing and the positive affect it produces were sufficient to evoke judgments of beauty, we would expect exaggeratedly plain stimuli to be considered the most beautiful. But that is not generally so; the well-known U-shaped relationship between complexity and beauty still has to be dealt with. Reber et al. are aware of this explanatory weak-spot and attempt to address it with Norbert Schwarz and Gerald Clore's "feelings-as-information" theory of affect (Schwarz & Clore, 1983). The feelings-as-information theory maintains that people utilize their affective states as sources of information when making evaluations, particularly when those feelings seem relevant: if the source of those feelings appears unrelated to what is being evaluated, said feelings are often discounted. In support of this hypothesis, Schwarz & Clore observed that on dreary days, subjects' weather-driven negative moods often shift their quality-of-life evaluations downward, i.e. on rainy days people are more likely to report dissatisfaction with their lives because they *misattribute* their bad moods to the state of their lives and not the dreary weather. Strikingly, however, that

influence is eliminated when interviewers draw the subject's attention to the true source of her melancholy by first inquiring about the weather. Citing this sort of finding, Reber et al. claim that the average person experiences distaste for extremely fluent, simple stimuli because the *source* of fluency/positive affect is discovered and then discounted. In other words, when simplicity makes it apparent that one's pleasure is merely the result of e.g. high repetition or obvious figure-ground contrast, we automatically discount that fluency and any pleasure it produces along with it.

But does this strategy really make sense of the data? I think not. There are a number of serious problems with Reber et al.'s attempt to explain the low-complexity/low-pleasure end of the U-shaped function between complexity and pleasure. First, there is no reason to think that the brute simplicity of a perceptual stimulus is enough to make perceivers more *aware* (either consciously or unconsciously) of the source of their affective states. Notice, for instance, that subjects often do not correct for misattribution of affect even in scenarios where the source of one's irrelevant affect is *glaringly obvious*. Schnall et al., for instance, found that clearly disgusting environments—e.g. a room containing a transparent plastic cup with the dried-up remnants of a smoothie, a chewed pen, and a trashcan over-flowing with greasy pizza boxes and dirty-looking tissues—still had an unconscious impact on how participants judged unrelated moral vignettes; in these studies, subjects found otherwise neutral stories to be more morally vexing when appraised in repulsive conditions (Schnall, Haidt, Clore, & Jordan, 2008). As the previous study by Schwarz & Clore would indicate, is often not until subjects

actively control for these appraisals via top-down reflection that misattribution is stemmed. That is to say, one usually must explicitly focus on e.g. the gloominess of the day or the dirtiness of the room to discount any unrelated displeasure that those conditions evoke. Although subjects questioned in aesthetic preference studies are rarely explicitly asked to provide justifications for why they dislike simple stimuli, it seems unlikely that they actively mitigate their initial feelings of approval after realizing they were responding to the "wrong" perceptual properties.[7]

Second, even if mere simplicity made us aware of the sources of our fluency-driven pleasure (e.g. repetition, familiarity, complexity, novelty, etc.), it is theoretically incoherent to suggest that our affective responses to those perceptual properties would be automatically discounted as *irrelevant* to the task of evaluating beauty or aesthetic preference. Here, the analogy to Schwarz & Clore's inclement-weather case seems to break down entirely. Our emotional responses to gloomy weather are quite clearly irrelevant to overall judgments of one's quality of life and so reasonably will be dismissed on reflection. But, whether an aesthetic object exhibits e.g. repetition, familiarity, complexity, novelty etc. seems *entirely relevant* to the evaluation of its beauty and aesthetic goodness. After all, the question of beauty and aesthetic pleasure is, in essence, a question about perceptual features, the combination thereof, and how they make us feel. If pleasure derived from the aforementioned aesthetic elements should be discounted as irrelevant to this

[7] Of course, I am not claiming that *all* discounting of irrelevant affect must be explicit: despite evidence of frequent misattribution of emotion, human beings are clearly capable of appropriately compartmentalizing their emotional appraisals. However, research into the misattribution of affect makes it clear that often such discounting does not arise without conscious reflection.

endeavor, then so too should form, shape, color, and every other perceptual property that can be processed either fluently or disfluently—but then, what is being judged as beautiful?

Third, while the feelings-as-information theory of affect does predict that we will discount affective appraisals upon discovering that they come from an irrelevant source, that effect is significantly stronger in the case of irrelevant *negative* emotions. That is to say, we are more likely to explain away or dismiss unrelated bad feelings than good ones (Schwarz, 2010). In Schwarz & Clore' inclement weather study, for instance, it was noted that the discounting effect was not obtained under sunny weather conditions. Even when subjects noted that the weather was enjoyable, they continued to evaluate their lives as better than average, suggesting that we are more inclined to dismiss sources of negative affect than positive. The problem is, if the fluent processing of perceptual stimuli is an intrinsically *pleasurable* experience, as Reber and his colleagues plausibly suggest, then overly simple stimuli would elicit positive appraisals that we would be *much less likely to dismiss* without significant incentive.

This is not the only empirical data point or observation that the processing-fluency view struggles to elucidate. The second criticism that can be levied against Reber et al.'s stance also involves the aforementioned U-shaped function, though for reasons that are quite distinct from the last. While it is important to explain our general preferences for perceptual properties that are e.g. neither too complex nor to simple, too familiar nor too novel, etc., it is also crucial to acknowledge the fact that these trends will apply "on the ground" in ways that are extremely flexible and

context-dependent. Unfortunately, reducing aesthetic pleasure only to the positive hedonic feedback of fluent processing makes that flexibility and context-sensitivity rather difficult to explain.

Consider the finding that art experts and novices show very different ranges of distribution along our inverted U-shaped complexity-beauty curve. While individuals with minimal art experience tend to prefer objectively simpler perceptual stimuli, individuals with art training appear to gravitate towards more complex and asymmetric aesthetic elements (e.g., McWhinnie, 1968).[8] To explain this fact, Reber et al. suggest that the art expert enjoys challenging stimuli because she has become skilled at effortlessly processing it; by seeking out repeat encounters with hard-to-process features, she learns to fluently process stimuli that are higher on the complexity scale and thus enjoys those stimuli more. At first glance, this explanation seems reasonable enough. After all, their view does predict that learning and competence will allow the expert to enjoy more complex perceptual features than the novice—and this is borne out by the data. On reflection, however, it becomes clear that this account is missing something crucial. For, notice that this explanation amounts to claiming by fiat that the art expert must *only* enjoy complex stimuli because it is *as easy* for her to process as it is for the novice to process simpler stimuli. While this might at times be the case, such a model does not make sense of the observation that people often report having a stronger sense of pleasure when processing inputs that initially challenge or continue to challenge

[8] Depending on one's level of art education, the inverted U-shape is shifted up or down, so to speak.

them. If we got maximal pleasure from easy fluency, why would we welcome grappling with stimuli that are ambiguous, complex, or novel in the first place?

Granted, Reber et al. are willing to concede that factors such as *cognitive* processing fluency may influence our ascriptions of beauty and sway our preferences, especially when it comes to the appraisal of artworks. Conceptual fluency reflects an individual's sense of ease when carrying out higher-level processes involving stimulus meaning and its relation to other semantic knowledge structures (Reber, Wurtz, & Zimmermann, 2004). Reber contends that in instances of art appreciation, an individual may well experience conceptual fluency with respect to e.g. artistic intent, identification of a style or genre, engaging with dramatic content, etc., even in the absence of explicitly *perceptual* processing ease (Reber, 2012). While this might account for our willingness to tolerate effortful perceptual processing at times, it casts our motivations in an unnecessarily metacognitive light. Surely there are times when an individual struggles to process stimuli for the mere sake of having that perceptual stimuli "click" or "make sense;" why we do that is mysterious on Reber et al.'s view.

More generally, the current processing-fluency theory seems to discount the impressive flexibility with which both art experts and laypeople alike are able to shift their appreciative stances to enjoy a broad range of complexity. As articulated, the view suggests that experts and novices are somewhat "locked in" to their patterns of aesthetic appreciation. If aesthetic pleasure and judgments of beauty correspond directly with *optimally high levels of processing fluency* (i.e. the point at which processing is as fluent as possible without being so obvious as to trigger

discounting) then Reber et al.'s model should also predict that individuals consistently prefer aesthetic objects that provide the *same range* of fluency processing. The problem is, common experience tells us that this prediction is unlikely to be confirmed. If affordance of fluent processing were the sole predictor of preference, then one individual art-lover with established processing capacities would not feel equally strong aesthetic pleasure in the presence of e.g. an intricate hybrid tea rose and a sleek arum lily. And yet, it seems quite possible for a single person to do just that. Indeed, the fact that people are capable of flexibly appreciating a wide range of perceptual stimuli with varying degrees of fluency affordance is a delightful and important-to-capture feature of aesthetic experience.

IV. DISFLUENCY, NEGATIVE AFFECT, & INTEREST

In light of these objections, it should be clear that the processing-fluency view has some significant inadequacies. As we have seen, the theory (i) does not currently deliver a robust explanation for why simple, very fluently processed stimuli are frequently unappreciated by the average observer; it also (ii) ignores the highly flexible and context-sensitive manner in which individuals can actively shift themselves into an appreciative aesthetic stance. I believe, however, that these explanatory gaps exist not because the processing-fluency view is wrong, but because Reber et al. fail to consider the crucial roles of disfluent processing, negative affect, and interest in aesthetic experience. They are missing the other half of the puzzle, so to speak. In his original arousal-potential theory, Berlyne astutely noted that our engagement with aesthetic objects entails an intricate interaction

between *interest* on the one hand and *pleasingness* on the other (in adapting his schema to the processing-fluency view, *pleasingness* can easily be understood as *perceptual-fluency*). So, what might we make of the interplay between *interest* and *fluent perceptual processing*? How might this new construct allow us to explain what Reber et al.'s view does not?

To answer this question, we should briefly consider the affective, motivational, and cognitive structure of interest. At the time of Berlyne's writing, virtually no one was writing on interest; since then, many researchers and theorists have contributed to a growing body of literature on that topic. While there is controversy today over how, exactly, we should think of psychological interest (for a review, see Silva, 2006), it is remarkable to note that many basic tenets of Berlyne's view are still largely corroborated by contemporary characterizations. The operating principle of Berlyne's view is that many organisms are motivated by what he called the *reduction of uncertainty*. In his arousal-potential theory, he argued that as the "uncertainty" or complexity of stimuli in an environment increases, an organism will enter into an aroused affective state that manifests itself as an exploratory impulse—e.g. it triggers behaviors such as foraging, tasting, seeking, smelling, uncovering, etc. In short, the animal becomes *interested* in attending to whatever stimuli is being perceived as novel or "uncertain." After engaging with the perturbingly "uncertain" stimuli for a while, the organism will be motivated to reduce that e.g. uncertainty, novelty, complexity, etc. into more easily processed, positively-appraised patterns.

In cognitive psychology, it is now a common assumption that most organisms have evolved basic exploratory states such as, e.g. “interest” (Izard, 1977; Tomkins, 1984), “seeking” (Panksepp; 1998), and “curiosity” (Gruber et al., 2014; Blanchard et al., 2015; Kidd & Hayden, 2015). And there is good reason to believe that these basic states are intrinsically motivating not only for humans, but across the animal kingdom. Macaque monkeys, for instance, show a clear preference for receiving information even when it has no impact on their reward rates; in fact, some primates even show a willingness to forgo a portion of their reward just to receive advance information about whether or not it will be delivered (Blanchard et al., 2015; Blanchard & Bromberg-Martin, 2015). Strikingly similar results have been found in research on pigeons, which reliably elect to receive less food when the choice is accompanied by information about the timing of that feeding—they prefer these “informative” feeding options even when the content of that information is negatively appraised, e.g. when they learn that their feeding will be postponed for longer than usual (Fortes, Vasconcelos, & Machado 2016). Presumably, if an animal is willing to sacrifice prized items such as food or water for the purpose of merely gathering new information of interest, then the process of resolving uncertainty is itself intrinsically rewarding, or hedonically marked.

Notice that this initial observation already gives us one possible avenue through which to explain what Reber et al. cannot with respect to the average perceiver’s distaste for extremely simple stimuli. If hedonic feedback produced by highly fluent processing were the *sole* source of affect driving our judgments of beauty, then we might indeed find exaggeratedly simple stimuli to be the most

beautiful or preferable. In line with the feelings-as-information view, however, I want to argue that the level of *interest* that a stimulus evokes also serves as an important cue to the observer about an object's pleasantness. In short, feelings of interest and processing fluency *both* influence our affective states and subsequent aesthetic experiences. It seems reasonable to posit that while high perceptual fluency may well be hedonically marked, the *strength* of the positive affect that it produces and the *intensity* of that appraisal seem to rely on how interesting and attention-grabbing we find the object in question. After all, when things are going well, one's perceptual system is constantly inundated with perfectly benign and perfectly fluent processing: we perceive a never-ending procession of fluent stimuli e.g. familiar people, cars, trees, lampposts, etc. yet, we do not walk around in a constant stupor over the aesthetic grandeur of it all. Rather, it isn't until we encounter perceptual stimuli *of particular interest* that that fluency seems to give rise to genuine aesthetic pleasure. In the case of, say, a simple outline of a circle (or any other austere stimulus), interest has virtually nothing to cling to.

If being the target of one's interest has so great an impact on the intensity and quality of our aesthetic judgments, we would do well to interrogate how stimuli come to hold our interest in the first place. One plausible answer to this question comes from Alison Gopnik's work on the role of negative and positive affect in human theory-formation and explanation (1998). On her view, affective phenomenology motivates our efforts to form theories about the world and marks the successful creation of "coherent" representational structures. As Gopnik describes it, theory-building is motivated by "hmm" moments (negative feelings of

anxiety, frustration, agitation, etc.) that arise when a system lacks relevant e.g. causal, spatial, conceptual, etc. representations; "aha" moments, by contrast, are marked by the positive feelings we get when "consistently" or "coherently" mapped representations are established. To cash this out in terms that will be relevant here, Gopnik's view highlights the fact that an initial experience of *dis*fluency (and the negative affect that it brings) is often required to get interest and motivation off the ground. Put simply, without the "hmm" moment, there can be no "aha!"[9]

Some empirically-based research in cognitive psychology seems to confirm the intuition that perceptual disfluency and resulting negative affect motivate and sustain interest. Brief episodes of negative affect associated with high-arousal emotions like confusion and frustration, for instance, appear to have an overall positive effect on problem-solving (D'Mello et al, 2009). It has also been found that participants who are primed with negative emotions not only experience enhanced motivation to finish challenging tasks (Liew & Tan, 2016), they also become more focused on seeking and using novel sources of information when solving problems (Spering et al, 2005).

The fact that processing disfluency and negative affect bear directly on our states of interest once again goes a long way towards explaining what Reber et al.'s view does not. One of the weak points of the original processing-fluency view is its

[9] To be clear, it is worth adding at this juncture that a third, prior appraisal likely stands as a necessary condition for interest to take root in the first place. In order for e.g. disfluency, uncertainty, complexity, and so on to motivate information-seeking, a subject must first *care about* or *potentially value* the source of that disfluency. The Macaque monkey, for instance, might be compelled to sacrifice rewards to clear up uncertainty about its next meal, but only because it values food. It would likely not give up rewards to complete its representational map of, say, where its human handler goes after work. It is because we care about things that disfluency matters and the displeasure of that disfluency becomes motivating.

inability to explain how we can experience aesthetic pleasure when engaging with stimuli that are subjectively disfluent or perceptually challenging. On an interest-fluency view such as the one I am advocating here, however, we would fully expect individuals to feel intrinsically motivated by the presence of especially effortful processing. In a manner of speaking, it is that disfluency and subsequent negative affect which signal to the viewer that there are worthwhile problems to be solved or ambiguity to be resolved. In fact, a recent study from Leknes and her colleagues might help us explain why the feelings of pleasure we experience upon resolving e.g. perceptual uncertainty are even *more* gratifying than many straightforward cases of fluent processing. While fluently processing stimuli may produce positive affect, Leknes et al.' study shows that subjective experiences of negative and positive affect can be pushed to the extreme by a contrast effect, i.e. fluency should be all the more pleasant when following a period of contrastive disfluency (Leknes, 2013).

Recently, theorists have turned their attention to the notion that exploration-motivating states such as interest are a kind of sui generis mental attitude of questioning (Carruthers, 2018.; Friedman, 2013). On such views, questioning attitudes are characterized as affective states that motivate one not just to seek information, but to seek information amending a specific state of ignorance, e.g. ignorance about *where I am*, *what that is*, *when something will happen*, etc. Importantly, these questioning attitudes do not need to be consciously articulated nor metacognitive to guide our behavior; presumably, much of the seeking and exploratory behavior we see across species is driven by questioning attitudes whose

content concerns the mere acquisition of propositional knowledge, e.g. what an object is, where it is, etc.[10] As this affectively-driven questioning attitude combines with more sophisticated reasoning capacities, we approach the kind of propositionally complex information processing that gives rise to conscious questioning and explaining in humans.

Given this understanding of primitive questioning attitudes, it would appear that a subject's level of interest in a stimulus not only has an impact on the *strength* of our aesthetic judgments, it also helps to focus our perceptual processes on various aspects of the stimulus that evokes our interest. It is through this ability to frame our aesthetic projects and focus our attention on different elements of a perceptual experience that we are able to appreciate a wide variety of stimuli regardless of the perceptual fluency that it affords. Indeed, one of the contributions that literacy in art theory makes to the aesthetic process is that it gives us a wider range of questions—both explicit and implicit—from which to draw.

V. CONCLUSION

While the processing-fluency view correctly focuses on hedonically marked, effortless processing as an important source of aesthetic pleasure, focusing *only* on fluent processing draws our attention away from an equally crucial aspect of aesthetic appreciation. As I have endeavored to show here, perceptual disfluency

[10] That said, a remarkable number of animal species (e.g. brown rats, crows, ravens, chimpanzees, etc.) do exhibit reasoning capacities and behaviors that are reflective of more robust understanding (Sawa, 2009). These animals are motivated to seek out information not only about what is in the world, but—at least to a limited extent—how and why things operate as they do.

and the negative affect that it produces also have a noteworthy role to play in our judgments of beauty. By piquing our interest, disfluency and negative affect have the power to motivate aesthetic engagement, amplify the pleasure we feel when resolving disfluency, and sustain the questioning attitudes that often make processing stimuli desirable in the first place. With this more elaborate understanding of the interaction between positively and negatively marked affective appraisals in place, the processing fluency view is in a much better position to stand up against criticism. Specifically, the view can now provide a more rigorous explanation of the inverted U-shaped relationship between complexity and aesthetic preference, and it can account for the immense variability of aesthetic judgements that individuals make on a daily basis. There is, of course, more work to be done to develop and test this view, but for now it should be clear that *displeasure* may play as fundamental a role in our ascriptions of beauty as *pleasure*.

CHAPTER 4

UPDATING THOUGHT THEORY: EMOTION AND THE NON-PARADOX OF FICTION

I. INTRODUCTION

The so-called "paradox of fiction" has been churning up controversy in philosophy of art for well over forty years now. The paradox, as originally stated by Colin Radford, is built upon three *prima faci*e plausible premises (Radford, 1975). The first premise is that (1) genuine and rational emotions require accompanying 'existence beliefs'—beliefs about the world and how that world might affect us. So, the story goes, when I see a bear in the woods at dusk, I am really and rationally afraid because I *believe* that I am actually in danger, e.g. I believe that *the bear might eat me!* The trouble is, if rational emotional states can only be predicated on beliefs about what exists, then our apparently emotional reactions to works of fiction look rather paradoxical. After all, when we imaginatively engage with fictions, (2) we do not believe those fictional characters or events to be real: at every turn we demonstrate through both action and conscious avowal that they are works of fiction. And yet, it also seems clear to most that (3) fictions do make us feel *something like* real and rational emotions. To borrow just a few popular examples from the debate, we cringe as Othello stalks towards Desdemona on the stage, we weep over the tortured fate of Anna Karenina, and gruesome monsters on the silver screen make us tremble in our seats. So, what are we to make of all of that cringing, weeping, and trembling?[1]

[1] Put formally, the paradox of fiction is generally expressed as follows:

Radford himself eventually concluded that our emotional reactions to fiction are altogether *irrational*, a claim that has proven less than satisfying to many theorists. At some point in time, each of Radford's three premises has come under strategic attack, leading philosophers to a number of surprising and often equally polemic conclusions. Kendall Walton's Pretend Theory, for instance, famously contends that since beliefs are a necessary component of genuine emotion, fictive scenarios can only evoke 'pretend' or *quasi*-emotional states—such reactions may be rational, but they are not real emotions (Walton,1978). Taking a different approach, proponents of Illusion Theory such as Alan Paskow propose that our responses to works of fiction are unproblematic because we *do* to some extent form what he calls 'illusory beliefs' about fictional events and characters (Paskow, 2004). And in opposition to these views, Thought Theorists such as Peter Lamarque and Noël Carroll attempt to reject the idea that emotions require anything like belief states about the world. On their account, we need only to represent (Lamarque, 1981) or 'entertain the thought of" (Carroll, 1990) evocative scenarios to experience genuine emotion, no existence-beliefs required.

At first glance, it would look as though every possible solution to the paradox has been explored. And yet, despite all of the careful and clever debate surrounding this topic, no single strategy has definitively put the puzzle to rest. One reason for this ongoing controversy, I propose, is that the aforementioned strategies are driven (at least in part) by intuition and fiat. It is standard practice for scholars

(1) To experience e.g. fear, one must believe that one is e.g., in real danger.
(2) When one knowingly engages with fiction, one does not believe that one is e.g., in real danger.
(3) When one engages with fiction, one experiences e.g. fear.

to endorse one or more of these premises as 'obvious,' and imprecise terminology in the literature abounds. Examples of this hand-waving are evident in all three of the well-known strategies aimed at resolving the paradox of fiction. Walton, for instance, deems the first premise (1) of the paradox to be 'common sense,' Paskow provides an extremely loose and contentious definition of 'belief' in his disavowal of premise (2), and Thought Theorists start from the third premise (3) by asserting that genuine emotional reactions to fiction *simply are* a commonplace occurrence. Because disagreement over the paradox of fiction can easily be chalked up to these differences in prior assumptions, the present-day debate is at something of an impasse.

What could break this standstill? Surprisingly, although the paradox of fiction was inspired by early cognitivist work on the ontology and structure of emotions, very little attention has been paid to contemporary scientific advances in that very domain. Since this puzzle first grabbed our attention in the early 1980's, theories of emotion in philosophy of mind, cognitive psychology, and neuroscience have changed dramatically. In that time, we have explored the neurological underpinnings of affect, discovered much about its expansive role in everyday cognition, developed new models of emotional processing, and theorized about its evolutionary function. These advances have a great deal to contribute to our understanding of the relationship between emotion, belief, and fiction—yet, they are rarely utilized to evaluate theorists' treatments of Radford's paradox.

In what follows, I aim to support and further develop Lamarque and Carroll's Thought Theory along empirical lines. This project is valuable for two

reasons. First, it is reasonable (as a criterion of endorsement) to want our best philosophical theories of emotion and art to align with current models of affect across the disciplines. I intend to show that only Thought Theory—with its rejection of the premise that emotions are predicated on existence-beliefs—can align in this way. Second, as I will explain, while few formal attacks have been levied directly against Thought Theory, its current lack of empirical support ultimately leaves the view open to doubt. By elaborating on Thought Theory with an appeal to recent findings on the cognitive structure and function of affect, I hope to more definitively resolve the paradox of fiction in favor of Lamarque and Carroll's view.

In section II, I will briefly review some of the thought-experiments and arguments often used to support the notion that (1) "existence beliefs" are a necessary component of emotion. Of particular interest will be the work of cognitivist philosophers Radford and Walton who, upon embracing this first premise, problematically conclude that our reactions to fiction are either *not instances of real emotion* or are *real but practically and theoretically irrational*, respectively. From there, I will attempt to fortify Thought Theorists' rebuttals to such lines of thinking. In section III, I will explore a generalized empirical account of emotion, focusing on the role of emotion in mind-wandering, counterfactual thinking, and prospection. As I will argue, the fact that we routinely recruit the affective system to engage in these instances of hypothetical thinking shows that we are quite capable of having real emotional responses to scenarios that we do not believe to be true. With the genuine status of emotional reactions to fiction safeguarded, in section IV I will argue contra Radford that these responses appear

both practically and theoretically rational (in the way that is appropriate to emotion rather than belief) in light of the evolutionary function of the affective system.

II. 'EXISTENCE BELIEFS' AND QUASI OR IRRATIONAL EMOTIONS

The paradox of fiction starts with the premise that emotional states are predicated on 'existence-beliefs' about the world. That is, it claims that for a scenario to really and truly sway us emotionally, we must *believe* that scenario to be real. This idea has been endorsed by a number of philosophers of art over the years, and its popularity has had a profound impact on the strategies devised for solving Radford's paradox. But, why think that existence beliefs are a necessary component of real or rational emotions in the first place?

The existence-belief criterion built into premise (1) owes its beginnings to early cognitivist theories of emotion. Around the time that Radford first published his now famous puzzle, cognitivist philosophers such as Robert Solomon began to explicitly emphasize the idea that emotions require propositional content (Solomon, 1988). That is to say, when we are e.g. afraid, that fear is informed by judgments of context and aimed at intentional objects—the fear is *about* something. In the case of our chance encounter with a bear, Solomon would maintain that while physiology may contribute to the richness of my emotion, it is fundamentally the *belief* or judgment that I am in danger that constitutes fear itself. This kind of cognitivist theory has some intuitive pull. As others have pointed out, there seems to be a clear causal relationship between our propositional attitudes and our

emotional states.[2] If, for instance, a last beam of light were to pierce the dusk and reveal our terrifying bear to be nothing more than an imposing statue, quite a different emotional picture would unfold. Given this new information, my belief would change (I would no longer judge that I am in danger) and my emotional state would presumably change along with it (perhaps to one of relief or amusement at my own readiness to flee).

Advocates of cognitivism have used these cases to posit that emotions should be identified strictly with the contents of one's beliefs; if no relevant belief or judgment is held, then no real emotion is instantiated.[3] As I will discuss in later sections, such a *purely* cognitivist model of emotion is no longer widely endorsed by scholars of affect. At the height of its popularity, however, this model inspired a number of well-known attempts at resolving the paradox of fiction. Presumably, it is with these cognitivist claims in mind that Walton proclaims, 'it seems a principle of common sense…that fear must be accompanied by, or must involve, a belief that one is in danger' (1978, p. 7). Radford similarly reiterates the postulate while defending his original statement of the paradox, claiming that 'I cannot be agitated or concerned for my safety, or the safety of others, by something I do not believe to be a possible danger to myself or others' (1977, p. 210).

So, what are the repercussions of granting this strong—albeit flawed—initial premise? In articulating his Pretend Theory, Walton eventually concludes

[2] Most notably, see Jerome Neu (2000) and Martha Nussbaum (2001).
[3] This should not be confused with the claim that *all affective* states require propositional content. Even on a strong cognitivist view, one could experience a change in affect qua e.g. moods or states of pain/pleasure without forming propositional beliefs. The existence-belief criterion is proposed specifically for the elicitation of emotions.

that knowingly engaging with fictions can at most elicit *quasi*-emotions, or mental states that are *not the same* as our emotional reactions to the real world. Crucial to his argument is the idea that genuine emotional states and their constitutive beliefs must reliably issue in action. So, Walton argues, when a moviegoer named Charles trembles at the appearance of 'the Green Slime Monster' on a movie screen, we know that a) Charles does not *really* believe he is in danger and hence b) Charles is not *really* afraid, because he does nothing to protect himself. As Walton astutely observes, audience members like Charles rarely *do* anything: he does not try to vanquish the monster on the screen, he does not run for his life, and he does not selflessly attempt to protect the children seated next to him. Although Charles may exhibit a few of the telltale physiological makers of emotion, such as trembling, his trembling is not accompanied by all of the other behavioral markers of 'true' fear because he lacks the requisite belief states. Walton's strong cognitivist leanings compel him to conclude that Charles is merely in a state of 'pretend' or *quasi*-fear; and since Charles is merely *playing at* emoting and is not in the grips of genuine, belief-based fear, Charles' response to the terrible monster looks more than anything like a rational game of make-believe.

While Radford is willing to grant that our responses to fiction are instances of *real* emotion (e.g. fear), he endorses cognitivism to claim that existence-beliefs are at least a requirement for *rational* emotions. Recalling our working example, imagine that I continue to carry on as if my life were in peril even after I realize that I am merely in the presence of a bear statue. According to Radford, being afraid of the statue when I know it can do me no harm is analogous to feeling afraid for

fictional characters: I know that there is no real danger, and so to emote as if there were an actual threat is to be *irrational* (1982). Radford does not elaborate on the *sense* in which emotional responses to works of fiction are 'irrational,' but as Richard Joyce has insightfully remarked, Radford's charge likely arises "from the dictates of two distinct normative frameworks" (2000, p. 216). On the one hand, Radford seems to take issue with our emotional responses to fiction because they are not *instrumentally rational*, i.e., if the function of fear is to e.g. promote safety, then being afraid while believing myself to be safe is of no *practical use.* On the other hand, Radford also seems to object that emotional reactions to fiction defy *theoretical rationality*; that is, he thinks that emotions not built on existence-beliefs are plainly ill-founded, i.e. they are not supported by the proper evidence. As Radford himself puts it, 'if I am frightened by a horror movie…it is not only helpful but correct to say to myself "Don't be silly!"…there is nothing to be frightened *of.*' So, even if we could benefit practically from engaging with fictions, emotional reactions to fictive entities would still look theoretically problematic or 'incoherent' on Radford's account (1982).

III. THOUGHT THEORY & REAL EMOTIONS

Many theorists have objected to the idea that our emotional reactions to fiction are *different* from the emotional states we experience when dealing with the real world. One of the most promising lines of attack against Walton's claim was developed by Lamarque in his account of Thought Theory. The Thought Theorist's strategy for dealing with paradox of fiction has generally been to flip the puzzle on its head:

the *real or genuine* status of our emotional reactions to fiction is granted as introspectively self-evident, and the task then becomes to explain the how these emotions are possible (Lamarque, 294). The explanation, for Lamarque, is that while fear in the real world is *often* inspired and directed by beliefs, there is no principled reason to think that *all* emotions are predicated on such existence-beliefs. He suggests, instead, that when we engage with works of fiction, we are able to respond emotionally not because we believe in the existence of e.g. Desdemona or the Green Slime Monster, but because we token *thoughts* or mental *representations* of those entities and the fictional worlds they inhabit.

The key point of Lamarque's theory is that thought-contents or representations provide us with mental descriptions (often in the form of propositional content) that do not get judged automatically as veridical or not. Take, for example, the thought-content *the bear is going to eat me!* Lamarque seems to suggest that when we entertain this thought—while safe, far from the woods at dusk—the content itself need not strictly involve judgments of truth-value. In other words, I can *think* about a predatory bear with a voracious appetite, but those thoughts do not necessarily become the target of my *beliefs* about what is true or plausible. I can think about such things while remaining doxastically unmoved in one direction or another. Of course, if I am faced with an actual bear, this thought-content can and often does immediately become the propositional content of a truth judgment, e.g. *the bear is* (really) *going to eat me!* Importantly, however, this belief-state is a further psychological attitude that one can hold *in*

relation to the thought being entertained, it is not necessarily part of the thought itself.

With this distinction between representation and belief in place, Lamarque argues that the actual trigger of any emotion is the content of one's thoughts. When the Green Slime Monster turns towards Charles and creeps forward on the movie screen, Charles responds emotionally to the thought-content *the monster is going to get me!* without forming any existence-beliefs about the slime or its actual ability to harm him. Far from paradoxically being frightened by things that are not really there to harm him, Charles is scared *by* the cognitive event tokened in the act of representing the Green Slime Monster; in this case, the *thought* of an approaching threat evokes his fear[4]. Lamarque goes on to clarify that Charles is not afraid *of* his thoughts—he is not terrified at the prospect of thinking things (at least, not under usual circumstances). What Charles is afraid *of* at a conscious level and in actuality is the intentional object of his fear, the Green Slime Monster, which need not be real or judged as real. According to Lamarque, this fear arises in the same way that any other real-world emotion would, hence it is genuine.

Such is the Thought Theorist's approach to defending the real status of our emotions when partaking in works of fiction, and it is promising one. If we do not need to believe we are e.g. in danger to experience genuine fear, then the paradox dissolves. As it stands, however, a number of weak points in the theory leave it open to objections by Walton and anyone sympathetic to his cognitivist views. On

[4] As I will emphasize in the next section, the precise causal role that these representations might play in eliciting emotion likely varies from scenario to scenario. At this juncture, however, we need only note the general shape of Lamarque's view for later elaboration.

what theoretical basis, for instance, can the Thought Theorists claim that our emotional reactions to fiction are the same as those we have when dealing with real-world scenarios? Lamarque has certainly provided a model for how psychological attitudes may be attached to 'thoughts' independently of one another *in principle*. The mere conceptual possibility of emotion without existence-beliefs does not, however, settle the question of whether emotions are *in fact* elicited in the absence of belief. Unless emotions reduce to mere phenomenology—and as I will explain below, there is good reason to think that they do not—then the observation that we 'feel' emotional when thinking about both real bears and fictional monsters does not settle the matter. To resolve this question, we would need to give an account of what emotion is and look to see if it is instantiated in a way that is consistent with Lamarque's model. Without this step, Thought Theorists can do little to persuade anyone not already inclined to intuitively agree.

So, what is an emotion? At first glance, this question might strike readers as overly ambitious. The nature of emotion has inspired countless philosophical and scientific investigations; models of emotion differ significantly in their characterizations, and at present, no single view has acquired consensus. What we might call "categorical" theories of affect (e.g. Ekman, 1992; Panskepp 1998), for instance, place emphasis on the notion that we have innate, discrete operating systems for each 'basic' emotion. By contrast, so-called 'dimensional' models (e.g. Russell, 1980; Russell & Barrett, 1999; Barrett & Wager, 2006) suggest that all affective states are instantiated via the same neurophysiological systems, but differ from each other along dimensions such as e.g. valence and arousal, and are

thereafter socially constructed and internalized.[5] Still others maintain that we have some "affect program emotions" that function like natural kinds, while others are higher-order and socially constructed (Griffiths, 1997). I certainly do not presume to resolve the conflict between these views here, and fortunately I do not need to. Despite their many differences, these and other competing theories converge upon a number of important points; any conclusions about fiction and emotion that we might come to by appealing to those commonalities should be on firm ground, regardless of whether emotions are natural kinds or not.

One general insight granted by most contemporary theorists is that emotion involves (i) physiological, (ii) evaluative, (iii) motivational, and (iv) phenomenological components. Although the extent to which these components are necessary or sufficient for emotion is highly contended, few would deny that each has a role to play. Revisiting our terrifying bear-encounter, cases of e.g. fear seem to involve a number of changes at both the physical and psychological level. As the bear rears up menacingly on its hind legs, for instance, (i) my heart-rate climbs and my breath quickens; (ii) I appraise the bear as potentially bad or threatening; (iii) I suddenly feel a strong inclination to run, and experience an acute focusing of attention on the bear and any possible escape-routes; and (iv) many of these bodily and mental states may be introspectively accessible, i.e. I *feel* that I am afraid.

Such changes are, of course, driven by different cognitive mechanisms. Many of these (i) physiological and (iii) motivational shifts are the product of a heightened state of *arousal*, and are roughly controlled by the reticular activating

[5] For more on the debate between categorical and dimensional models of affect, see e.g. Zachar & Ellis (2012).

system in the brainstem (involved in sleep and attentiveness regulation), the autonomic nervous system (responsible for changes in involuntary motor functions) and the endocrine system (the release of e.g. adrenaline) (Pfaff, 2015). But, emotions also have a crucial (ii) evaluative component. When we experience any given emotion, that emotion will be *valenced* to some extent or another, i.e. it will carry with it either positive representations of "*seeming-goodness"* or negative representations of *"seeming-badness*." While changes in emotional arousal may at times be automatic, is clear that (i) physiological and (iii) motivational states often interact robustly with these (ii) valenced appraisals. After all, it is presumably the fact that I tag '*the bear might eat me!*' with bad-seeming valence that activates my heightened state of arousal and subsequent flight-or-fight response.[6]

While there is solid evidence that unique mechanisms exist for the processing of pleasant and unpleasant stimuli (Viinikainen et al., 2010), valence *in general* is encoded within a number of limbic areas, including the anterior insula, the ventromedial prefrontal cortex and the orbitofrontal cortex (Rainville et al., 1997; Peyron et al., 2000; Levy & Glimcher, 2012). These regions are consistently activated across a diverse range of affective states (Phan et al., 2002), and localized brain trauma or impairment appears to have a devastating impact on emotional processing (Damasio, 1994). Together, these systems contribute to the (iv) experiential quality of our fears, joys, and everything in between.

It is worth noting in this general characterization that emotion is just one of many affective states: what makes emotion different from, say, a mood, or pain and

[6] Were a world-renowned bear-trainer to find herself in the same scenario, the valence of her appraisal would likely differ, and her emotional experience would be quite unlike mine.

pleasure? Emotion appears to be unique in the sense that it is *about* something. As the cognitivists remarked long ago, emotions such as fear take specific targets. When changes to arousal and valence are attributed to a particular target or occurrence, we are said to be experiencing an emotion. I am afraid *of the bear*, angry *at myself* for going into the woods at dusk, and relieved *by the fact* that it was only a statue.[7] Notice, though, that the target and cause of one's emotional state need not be one and the same. Extensive research on the fungibility of affect has found that negative valence can often be caused by unpleasant stimuli and then attributed to an unrelated target. Reading moral vignettes while in a dirty room, for instance, has been shown to intensify feelings of moral outrage and anger. While the target of that anger is the moral vignettes (the characters therein are *bad-seeming*), the cause of that anger is presumably misattributed negative-valence—valence that is caused by appraising unpleasant conditions (Schnall et al, 2008; Li et al, 2007).[8]

In sum, the empirical literature gives us good reason to hold that emotions are comprised of changes in arousal and/or valence at the level of (i) physiology, (ii) evaluation, (iii) motivation, and (iv) phenomenology, which are then attributed

[7] It is interesting to note that when the object of one's altered arousal and valence is specifically a bodily sensation (say, the pinch of a needle), it is often classed as pain or pleasure. Changes to these networks accompanied by no specific target at all begin to look more like moods, e.g. when the emotion targets nothing in particular, the high arousal and negative valence of fear becomes a *mood* of anxiety.

[8] Assuming that we do in fact respond to fiction with genuine emotion, it is important to point out that when we e.g. watch Othello on the stage, the emotions that arise in us may be caused by a multitude of things. One could, for instance, feel sad because the drama reminds her of real suffering and real pain in the world. Or, one might have a general mood of anxiety that is given expression through the fiction. But, notice that such responses do not fall under the purview of the paradox of fiction: if an actor's fictional predicament triggers sadness, but the target and cause of that sadness is real suffering or one's own state, then that emotional response is predicated on an existence belief and it is no longer puzzling on a cognitivist reading.

to a target[9]. With this barebones account in place, it should be a relatively straightforward task to justify the Thought Theorist's foundational assumption that emotional reactions to fiction are real. To substantiate this premise, we would need to show that engaging with fiction and nonfiction elicits the same kind of output (i-iv), is driven by the same neurological mechanisms (the affective brain-networks described above), and is the result of processing the same kinds of cognitive content.

The first of these steps is clear enough—as Theory Theorists note, few would deny that we (iv) introspectively *feel* like we have emotional reactions to fiction. When prompted, most will avow that they are worried about Desdemona, saddened by Anna Karenina, and afraid of the Green Slime Monster. And it is also certainly the case that our reactions to fiction can have profound (ii) physiological impacts: just remember all of that crying, weeping, and trembling! Although it is true (as Walton noted) that our reactions to fiction and non-fiction may produce distinct outward behavioral patterns, that behavioral difference does not reveal a (iii) motivational one. In fact, the Thought Theorists may happily concede Walton's point that beliefs strongly impact our actions, but deny that overt action is constitutive of our emotions. On this reading, when Charles is in the movie theater, his emotional response (i-iv) includes motivation to e.g. run away, hide, etc., but his actions are *also* controlled and down-regulated by the belief that he is not *really*

[9] Though I will not pursue the topic at length here, I do not assume that emotional attribution needs to be conscious. One could, I believe, experience heightened arousal/negative valence and represent a target as bad-seeming without direct awareness of the target's newly acquired status. Representations of the target as bad-seeming could have behavioral repercussions (e.g. frowning when the target is in view) in the absence of any conscious awareness, hence the thought "I was angry at him but didn't even know it."

in danger. This does not, however, mean that the motivational states associated with fear were not activated[10].

Whether or not our reactions to fiction include (ii) valenced appraisal is an empirical question—one that lines up in general with the question of whether our emotional responses to fiction rely on the same general cognitive mechanisms as those that produce our emotions to everyday events. Fortunately, a number of studies have been conducted to explore the relationship between fiction and the affective arousal-valence network. In 2008, for instance, Jabbi et al. scanned subjects at varying points while they imagined disgusting gustatory experiences, tasted an unpleasant liquid, and viewed actors pretending to do the same. They found that under all of these conditions subjects displayed clear neurological markers of disgust. Specifically, brain scans showed increased activity in the anterior insula (a region strongly linked to representations of negative valence) whether participants were imagining the foul liquid or actually tasting it themselves. Similarly, in 2010 Bray et al. found that receiving real and imagined rewards both engage overlapping areas of the orbitofrontal cortex, the primary neurological network associated with valence. And in 2011, Wallentin et al. confirmed that listening to intense fictions leads to an increase in heart rate along with increased activation of the amygdala and other brain regions associated with emotional arousal.

These empirical studies, and others like them, go a long way towards showing that imaginative engagement with fiction produces genuine emotions.

[10] For more on the interaction between belief and emotion, see next section.

Reactions to the *imagined* Green Slime Monster and the *real* bear are churned out by the same cognitive processes and produce the same sort of phenomenology and targeting of objects. In one case, we know the target of our emotions to be imagined and in the other case we believe the target to be real. But should that matter? Why should we think, as Walton contends, that emotions require input in the form of existence-beliefs to count as genuine emotion?

Certainly, if we take our encounter with a bear in the woods at dusk to be the only archetype of true emotion, then Walton's objection poses a serious concern. As the cognitivists made clear long ago, existence-beliefs do factor strongly into our fearful attempts to flee real-life danger. Over the past several decades, however, it has become apparent that our emotional architecture is involved in responding to much more than just our perceived environments.[11] We now understand that the brain networks associated with affect play an extensive role in everyday modes of cognition that arguably do not require anything like belief-states as input. It has recently been shown, for instance that affect is crucial for "prospection" (Gilbert & Wilson, 2007), or our ability to weigh options and make decisions about future states that are as yet uncertain or counterfactual. When Benoit et al (2014) investigated brain activation during guided acts of prospection, they found that the ventromedial prefrontal cortex (associated with affective processing) was active when subjects were instructed to imagine both

[11] In the 90's, Richard Moran astutely observed that cognitivists often cherry-pick cases of "typical" emotion in favor of those that seem to require existence-beliefs (1994). Since that time, many scientific studies (some of which I will turn to here) have both confirmed the connection between emotion and counterfactual thinking, and demonstrated just how prevalent this kind of affectively-laden thinking is.

familiar/realistic and entirely novel/improbable future events. Benoit and colleagues hypothesized that this activation of valence and arousal networks underlies our ability to predict possible future affective states (i.e. would this make me happy or sad?) and guide any future-oriented decisions accordingly. And this is just one of several ways in which we affectively engage with imagined scenarios on a frequent basis through what is often called "mind-wandering" (Buckner, Andrews-Hanna & Schacter, 2008). When allowing our minds to wander from the tasks in front of us, for example, we also frequently engage imaginatively with the past, ruminating on e.g. what should have been said or done (but was not), what would have occurred had things happened differently, etc. And this kind of thinking is not a rare occurrence; as much as 50% of our waking hours are spent thinking about what we *would* have done, *could* have done, *should* have done, and *might* do differently in the future (Killingsworth & Gilbert, 2010).

These forms of mind-wandering all heavily recruit affective processing. But notice that in order for these modes of cognition to operate, they *cannot* rely exclusively (if at all) on belief-states as input. When I partake in prospection about what it would be like to go on a vacation to Spain, for instance, I surely do not need to *believe* in any reasonable sense that I *will* go to Spain in order to weigh the possibility affectively—the very purpose of affectively-laden prospection is to steer an open course! Similarly, when I obsess over what I should have said during last week's lecture, there is no sense in which I *believe* myself to have both bored my students to tears and enthralled them at the same time. The positive valence and arousal evoked by thinking of how well the class *could* have gone is stored in

memory and helps me make well-reasoned future decisions, but no existence-beliefs are required for that appraisal.

So, if affectively-driven counterfactual thinking about the world is not predicated on beliefs, then what does it take as input? To use Lamarque and Carroll's terminology, it would appear that these acts of imagination simply require thought. But, this needs some elaboration beyond mere 'mental representations,' or Fregean 'senses' or 'descriptions.' Although he does not offer much more than these terms, Lamarque does gesture in a promising direction when he remarks that 'thoughts can differ among themselves with respect to vividness' (295). Given that thought-content must be open to this variation in vividness, why not posit that thoughts are propositions which are sensorily-formatted—e.g. the processing of images from a movie screen, engaging with mental pictures formed while reading a book, auditory representations of speech, representations of imagined tactile or gustatory experience, etc.? The advantage of identifying thoughts with sensory-motor representations is that, first, such representations may (but do not necessarily) become the targets of attitudes such as belief or emotion. And second, sensorily-formatted content will engage the affective system to varying degrees of intensity based upon the level of detail expressed in that representation. As Metcalf and Mischel (1999) have discussed extensively, the same propositional content e.g. *the bear is going to eat me* can be affectively salient or not depending on the extent to which it triggers 'hotspots' or 'cold nodes.' If, for instance, that propositional content is sensorially rich (imagine, for instance, the smell of the bear's rancid breath as it lunges forward, its sharp yellowed teeth snapping at your throat), it will

likely engage the affective system more intensely that a sensorially sparse representation.[12]

By now, it should be clear that the cognitive mechanisms which drive our everyday emotional states can and very often do process input that is not accompanied by attitudes of belief or disbelief. This fact, in tandem with the finding that reactions to fiction are processed in the same affective brain-areas as nonfictional events should be enough to corroborate Thought Theorists' intuition: emotional reaction to fiction are *genuine*, and *real*. But, we are still left with another pressing question. Are these emotional reactions *rational*?

IV. THOUGHT THEORY AND RATIONAL EMOTIONS

As I mentioned at the outset, Radford takes issue with our emotional reactions to fiction on the basis that they are "incoherent." Arguing against Thought Theory, he suggests that if something like a movie monster is the target of Charles' fear, then his fear might be real, but it is irrational; it is irrational because what frightens Charles can neither harm nor interact with him in any way (1982b, 261-62). How are we to unpack this charge of irrationality?

If it is the *instrumental or practical* rationality of Charles' fear that Radford is concerned with, then he must be worried that emotional reactions to fiction work against (or fail to further) one or another of Charles' aims. But, what aim does fear of the fictional Green Slime Monster fail to achieve? As many have argued before

[12] To be fair, Carroll in particular elaborates on the notion that thoughts about 'Dracula' point to the Fregean 'sense' of the name or 'the congeries of properties attributed to the vampire in the novel' (85). But, I find that the Fregean framework does little to explain *how* thoughts come to evoke affect at the cognitive level. Sensory-formatting should help to fill in that gap.

me, it would appear that fiction's strong elicitation of genuine emotion allows us as readers, viewers, listeners, and audience members alike to reach a number of practically desirable ends. To name just as few, partaking in fiction emotionally has been shown to increase empathy (Bal & Veltkamp, 2013), such emotions are the anchor point of important social institutions around fiction (Carroll, 1990), and these emotions are fundamentally pleasurable—they are themselves a practical aim (see Joyce, 2000). And now, after looking at the role of the affective system in counterfactual thinking and prospection, it becomes even clearer that affectively-laden imaginative thought plays a crucial role in our fitness as a species. Even when they are not themselves plausible or realistic, fictions often provide us with potentially useful emotional reference points: although Charles may never need to know how he would handle a *monster per se*, his imaginative encounter with The Green Slime Monster leaves him more familiar with his own reactions to danger and how to cope with those novel feelings.[13] Practice makes perfect, as the saying goes. According to Thought Theorists such as Carroll, these emotional reactions—as long as they are not 'psychotic, or even neurotic, Fantasies'—do not interfere with our lives in a way that is *practically* irrational.[14]

Radford himself concedes this point, admitting—as an admirer of art and aesthetics—that in cases such as that of Anna Karenina, 'it might even be said to be irrational' to discourage emotional engagement with fiction (1982b). But

[13] As Dennett (1995) has commented in other contexts, this remarkable ability to imaginatively weigh our options 'permits our hypotheses to die in our stead.'
[14] Joyce (2000) builds on this idea by cleverly arguing that we partake in fictions with the aim of being emotionally swayed, and so we fulfill our instrumental aims when being moved by works of fiction.

Radford doesn't retract his charge of irrationality; weeping over Anna's fate must be irrational because it overthrows *some* aim. Which aim? In the many iterations of his complaint, Radford seems to betray a worry about the teleology or purpose of emotions *themselves*. By decrying the fact that we fear and mourn characters that do not exist and cannot interact with us, he seems to suggest that the *purpose* of emotion itself is to elicit action or behaviors that are salient in the real environment. Charitably, his concern seems to be that our responses to fiction are practically irrational because they frustrate their very own ends.

While there is strong evidence that emotions evolved in large part to spur evolutionarily advantageous behavior (see, e.g. Tooby & Cosmides, 2008), it would be a mistake to think that the teleology of affect is bound up exclusively with the elicitation of action or behavior *on its own*. This is precisely why belief—though conceptually distinct from affectively evocative thoughts—has such an important role to play in our emotional lives. To better understand the complex ways in which affective judgments of valence and judgments of veracity can inform our emotional states, we should turn our attention to contemporary models of affective appraisal.

To date, one of the most prominent and well-accepted theories of affect, Appraisal Theory, claims that beliefs are one of many propositional attitudes that factor into the generation of our affective states. Drawing on the merits of both cognitivism and non-cognitivism, Appraisal Theorists posit that emotions are driven by a combination of knee-jerk 'associative processing' or 'innate' sensory-motor responses (respectively, Kirby & Smith, 2000; Marsella & Gratch, 2009) in tandem with cognitive appraisals of context, beliefs, goals, etc. On this view, when

Charles sits in the movie theater as the monster oozes across the screen, he may very well feel immediate fear at the sight of the monster's cavernous, sickly maw and the sound of its deathly moan. The mere thought or sensorily-formatted representation of the creature's disgusting, 'hot' properties, though unaccompanied by belief, can be appraised for seeming goodness and seeming badness. Charles' appraisal in this case is, namely, that slime monsters are *bad*. Importantly, however, the affective appraisal *monsters = bad* is just one of many kinds of input being appraised by Charles: he also believes that *he is safe in the movie theater*, he judges that *those are impressive special effects*, his goal is to *see how the director resolves this tension*. While valenced appraisals of the Green Slime Monster can non-doxastically elicit fear, that negatively-valenced appraisal can then be assessed and mitigated in the wider context of Charles' aforementioned beliefs, thoughts, values and goals.[15]

If the cognitive mechanisms associated with affect have a practical purpose to be hampered, it is not to assess the veracity or probability of thought-content or how to go about interacting with our environments—that requires much more complicated input. The purpose of affect is simply (though not insignificantly) to assess value and draw attention to the targets of our emotional output. When we respond emotionally to fictions, we succeed to that end.

[15] This may go some way towards resolving another well-known paradox in aesthetics that is built around how we are able to derive pleasure from unpleasant states, or the 'paradox of tragedy.' The fact that the affective appraisal system responds simultaneously to multiple e.g. belief and thought contents makes sense of the recently confirmed finding that reading gut-wrenching fictions evokes both positive and negative valence (Altmann, Bohrn, Lubrich, Menninghaus, & Jacobs, 2012). In such cases, Appraisal Theory allows that we are able to respond to the *thought* of Anna Karenina's fate with genuine grief, while also appraising Tolstoy's masterful evocation of sadness as good-seeming.

But what about the *theoretical* rationality of our affective tendencies? In general terms, to be theoretically irrational is to be ill-founded or under-supported by the proper evidence. Usually, this standard is used to evaluate doxastic states—when both evidence and argumentation support the truth of a propositional belief, that belief is theoretically rational. It is not uncommon for there to be tension between our theoretically rational and instrumentally rational beliefs. I could have a great deal of evidence to support the belief that I am not my mother's favorite child, for example, but it may not be in my best interest psychologically to believe it. Over the span of many years and across multiple articles, Radford seems to suggest that emotions may exhibit this same tension: though their practical benefits are many, he holds that there is something plainly wrong about building emotions on representations of fiction—it is like building belief on contrary evidence.

In response to this criticism, Carroll states that perhaps 'it is just a fact about humans that they can be frightened of the *idea* [of fictional entities]…It is just the way we are built' (83). He goes on to argue that our emotional reactions to the Green Slime Monster are not 'silly,' because they are normatively appropriate; they are rational in the sense that they are prescribed. Despite the intuitive upshot of Carroll's defense, Thought Theory does not completely forestall Radford's claim that e.g. cringing for Desdemona is 'irrational and incoherent.' Just because emotional reactions to fiction are normatively accepted or common does not mean they are *theoretically* rational. Further, even if humans are endowed with the capacity to be frightened by thoughts, how does the universality of that capacity thwart the accusation of irrationality? If we do not need to *believe* fictions to emote

rationally, then what exactly is our cringing, weeping, and trembling *about* and how is it *theoretically* rational?

The problem with Radford's grievance, I argue, is that the standards of theoretical rationality which hold intuitively for beliefs plainly do not translate to cases of affective appraisal. As we have already seen, representations of valence are not in themselves representations of truth : they are measures of value, or seeming-goodness and seeming-badness (Levy & Glimcher, 2012). As a result, one might think that an emotional response is rationally grounded (that is to say, theoretically rational in the sense appropriate for emotions) to the extent that the evaluation contained in the response is an appropriate match for the representation whose appraisal provides the input for that response. On this view, fear at a ferociously-represented bear may then be rational or irrational based upon whether that representation includes elements that are good-seeming or bad-seeming, regardless of whether the bear is really there being perceived, or merely imagined.

As it stands, there are numerous theories detailing what might make one's emotional responses theoretically rational. For our purposes here, it is worth noting that some of these theories focus precisely on the notion of appropriateness or fitness just mentioned. Ruse and Wilson (1968), for instance, argue that the theoretical rationality of any given emotion is rooted in the biological functions of those emotions; happiness should target events that are potentially beneficial to the organism, sadness should target those objects and events that might involve potential loss, etc. In the case of our bear at dusk, fear is rational because it is directed towards a creature and situation that can *at least in principle* be dangerous.

While I am hesitant to apply this standard to emotions *simpliciter*—in part because I wish to remain impartial on the question of whether emotions are natural kinds—the underlying principle of these 'fitness' views speaks to what makes the *appraisal* component of emotion seem well founded or ill founded. While a person need not believe the cause of her e.g. fear to be real in order for that fear to be rational, she does need to have some evidence that the target of her fear could *hypothetically* be bad in some way.

Evidence for the fittingness of our appraisals can be both innately channeled and learned. Presumably, many affective valuations initially evolved and were passed on to confer some evolutionary advantage e.g. bodily damage will often generate negative valence/high arousal to result in pain, while sugary foods that provide life-sustaining energy will likely generate positive valence/high arousal, and result in pleasure (Berridge & Kringelbach, 2013). At the same time, individual experience and learned association should also factor strongly into whether an appraisal is appropriate or theoretically rational. For instance, the seeming-badness of our bear encounter seems rational or appropriate enough for the average hiker, but a life-long bear trainer and wildlife enthusiast would presumably be elated by the encounter. If both of these reactions seem rational, it is because their appraisals of the potential dangerousness of the very same perceptual representations are appropriate to each case. Given her past (in)experience with bears and an innate predisposition towards avoiding particularly large predatory animals, the hiker has good reason to appraise the bear as dangerous; the trainer, by contrast, has good evidence that bears can be harmless when handled properly, and so her enthusiasm

seems equally appropriate. Were these appraisals to occur in response to merely imagined or fictional contents, the fittingness criteria would hold just as well.

Admittedly, while I find this 'appropriateness' model to be persuasive, it is one of many views on the theoretical rationality of emotion.[16] My aim here is not, however, to hold up such 'fittingness' views as the only plausible answer to questions about the theoretical rationality of emotions. Rather, I want to highlight the fact that our criteria for assessing the theoretical rationality of belief need not—in fact, *should not*—apply in the case of valenced appraisals or emotion in general. To assume that emotion must be grounded in evidence of veridicality is simply to beg the question in favor of cognitivist's intuitions. Given that the affective system is routinely recruited for thinking about fictional or imagined events, we have every reason to maintain that appraisals can be theoretically rational or irrational based upon grounding in e.g. perceived value, not evidence of existence.

V. CONCLUSION

As should now be apparent, Lamarque and Carroll's work on Thought Theory went a long way towards defending the status of our emotional reactions to fiction as both real and rational. But, while the theory championed our intuitive assumptions about emotion and fiction, these theorists did little to motivate their founding premise that emotional reactions to fictions count as genuine emotions. By looking at the cognitive science of emotion and the role that affect plays in counterfactual mind-wandering and prospection, I hope to have demonstrated that emotional states

[16] For a more detailed look at this issue, see Deonna & Teroni (2012)

need not be predicated on anything like belief-states about the world. Further, by examining the function of affect, I hope to have shown that emotion is not susceptible to the same charges of theoretical irrationality as, say, belief. As I have indicated, researchers in philosophy of mind and the cognitive sciences still have many fascinating questions to address with respect to emotion, affect, and appraisal. In light of what we now know, however, the paradox of fiction is starting to look not so paradoxical after all.

REFERENCES

Altmann, U., Bohrn, I. C., Lubrich, O., Menninghaus, W., & Jacobs, A. M. (2012). The power of emotional valence—from cognitive to affective processes in reading. *Frontiers in Human Neuroscience*.

Armstrong, T., & Detweiler-Bedell, B. (2008). Beauty as an emotion: The exhilarating prospect of mastering a challenging world. *Review of General Psychology, 12,* 305.

Arnold, Magda B. (1960). *Emotion and Personality*, New York: Columbia University Press.

Avenanti A, Sirigu A, Aglioti SM. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*; 20 (11): 1018-22.

Bal, P.M., Veltkamp, M. (2013). How Does Fiction Reading Influence Empathy? An Experimental Investigation on the Role of Emotional Transportation. *PLOS One* 8(1) January 30, 2013

Baron-Cohen, Leslie & Frith. (1985). Does the autistic child have a 'theory of mind'? *Cognition* 21 (1): 37–46.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. In A. Whiten, Ed., *Natural theories of mind: Evolution, development, and simulation of everyday mind-reading*. 233-251. Cambridge, MA: Basil Blackwell.

Barrett, L. F., & Wager, T. D. (2006). The structure of emotion. Evidence from *Neuroimaging Studies. Current Directions in Psychological Science*, 15(2), 79–83.

Benoit, R.G., Szpunar, K.K, & Schacter, D.L. (2014). Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences of the United States of America,* 11:16550-16555.

Berlyne, D. E. (1974). *The New Experimental Aesthetics*. New York: John Wiley & Sons.

---. (1971). *Aesthetics and Psychobiology*. New York: Appleton-Century-Crofts.

Berridge, K. C., & Kringelbach, M. L. (2013). 'Neuroscience of affect: Brain mechanisms of pleasure and displeasure.' *Current Opinion in Neurobiology*, *23*(3), 294–303.

Blair, J, Sellars, C, Strickland I, Clark F, Williams A, Smith M & Jones, L. (1996). Theory of Mind in the Psychopath. *The Journal of Forensic Psychiatry.* 7:1, 15-25.

Blair J. (2012). Considering anger from a neuroscience perspective. *Wiley Interdisciplinary Review of Cognitive. Science.* 3, 65–74.

Bloom, P. (2016). *Against Empathy: The Case for Rational Compassion*, New York: Ecco.

Botvinick, M. et al. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *Neuroimage* 25: 312–319.

Bray, S., Shimojo, S., O'Doherty, J. (2010). Human Medial Orbitofrontal Cortex Is Recruited During Experience of Imagined and Real Rewards. *Journal of Neurophysiology*. 103; 5.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*(1), 1–38.

Buttelmann, David & Zmyj, Norbert & Daum, Moritz & Carpenter, Malinda. (2012). Selective Imitation of In-Group Over Out-Group Members in 14-Month-Old Infants. *Child Development*; 84(2):422-8

Byrne, R.W. (1998). Imitation: the contributions of priming and program-level copying. In *Intersubjective communication and emotion in early ontogeny* (ed. S. Braten), 228–244. Cambridge University Press.

Cabanac, M. (1992). Pleasure: the common currency. *Journal of Theoretical Biology*. 155, 173–200.

Carroll, N. (1990). *The Philosophy of Horror; or, Paradoxes of the Heart*. New York: Routledge.

Carruthers, P. (2018). Valence and Value. *Philosophy and Phenomenological Research*, 97, 658-80.

-- (2018b). Basic questions. *Mind & Language*, 33 (2018), 130-147.

--- (2006). Why Pretend? In S. Nichols (Ed.), The Architecture of Imagination. Oxford: Oxford University Press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76,* 893–910.

Child, I. L. (1962) Personal preferences as an expression of aesthetic sensitivity. *Journal of Personality*, 30: 496–512.

Csibra, F. (2007). Action Mirroring and Action Interpretation: An Alternative Account. in *Sensorimotor Foundations of Higher Cognition (Attention and Performance XII)*, ed. P. Haggard, Y. Rosetti and M. Kawato, 435–459. Oxford: Oxford University Press.

Damasio, R., (1994). *Descartes' error: Emotion, reason, and the human brain* (3rd ed.). (New York: Putnam Pub Group).

Davies M. and T. Stone (eds.), (1995). *Folk Psychology and Mental Simulation*, Oxford: Blackwell Publishers.

Decety J, Chen, C., Harenski, C., Kiehl, K.A. (2013). An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Frontier Human Neuroscience.*

DeLoache, J. & LoBue, V. (2009). The narrow fellow in the grass: Human infants associate snakes and fear. *Developmental Science.* 12. 201-7.

Dennett, D. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. (New York: Simon & Schuster).

Deonna, J. A. and Teroni, F., (2012) *The Emotions. A Philosophical Introduction.* London/New York: Routledge.

de Sousa, R (1987). *The Rationality of Emotion*, Cambridge, MA: MIT Press.

Destro, Boria, Pieraccini, Monti, Cossu, et al (2007) Impairment of action chains in autism and it possible role in intention understanding. Proc Natl Acad Sci U S A. 104(45):17825-30. Epub.

de Vignemont & Jacob, P. (2012). What Is It like to Feel Another's Pain? *Philosophy of Science*, 79 (2): 295-316.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., van Baaren, R. B. (2006). On making the right choice: The deliberation-without- attention effect. *Science,* 311, 1005–1007.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* 91: 176–180.

D'Mello, S.K., Person, N., & Lehman, B.A. (2009). Antecedent Consequent Relationships and Cyclical Patterns between Affective States and Problem Solving Outcomes. Proceedings of 14th International Conference on Artificial Intelligence in Education (Brighton, UK, July 6-10, 2009), 57-64.

Dolan, M., Fullam, R. (2004). Theory of mind and mentalizing ability in antisocial personality disorders with and without psychopathy. *Psychological Medicine* 34, 1093–1102.

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child development*, *82*(3), 793–811.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review.* 66 (3): 183–201.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*. New York: Oxford University Press 6 (3/4): 169–200.

Farkas, A. (2002). Prototypicality-effect in Surrealist paintings. *Empirical Studies of the Arts, 20,* 127–136.

Fecteau, S., Pascual-Leone, A., Théoret, H. (2008). Psychopathy and the mirror neuron system: preliminary findings from a non-psychiatric sample. *Psychiatry ResPsychiatry Res*; 160(2): 137-44.

Fortes, I., Vasconcelos, M., & Machado, A. (2016). Testing the boundaries of "paradoxical" predictions: Pigeons do disregard bad news. *Journal of Experimental Psychology: Animal Learning and Cognition, 42*(4), 336-346.

Friedman, J. (2013). Question-directed attitudes. *Philosophical Perspectives,* 27, 145-174

Friedman, O. & Leslie, A.M. (2004). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science*, 15, 547-552.

Frijda, N. H. (1994). Varieties of Affect: Emotions and Episodes, Moods and Sentiments, in Ekman & Davidson (eds.), *The Nature of Emotion: Fundamental Questions*, New York: Oxford University Press, 59–67.

Gallagher, H. L., Frith, C. D. (2003). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences 7 (2):* 77–83.

Geangu, E; Benga, O; Stahl, D; Striano, T. (2010). Contagious crying beyond the first days of life. *Infant Behavioral Development*. 33 (3): 279-88.

Gendler, Tamar Szabó, (2000), "The Puzzle of Imaginative Resistance", The Journal of Philosophy, 97(2): 55–81.

Gilbert, D. T., & Wilson, T. D. (2007). 'Prospection: Experiencing the future.' *Science*, *317* (5843), 1351–1354.

Goldman, A. I. (2006). *Simulating Minds*. Oxford, Oxford University Press.

Gopnik, A. (1998). Explanation as Orgasm. *Minds and Machines,* 8, 101-118.

Gordon, R. M. (1995). Simulation without introspection or inference from me to you. In Davies and T. Stone (eds.), *Mental simulation: Evaluations and Applications*. Oxford, Blackwell, 53–67.

Graf, L. K., Mayer, S. and Landwehr, J. R. (2018), Measuring Processing Fluency: One versus Five Items. J Consumer Psychology, 28: 393-411.

Griffiths, P. E., (1997), *What Emotions Really Are: The Problem of Psychological Categories*, (Chicago: University of Chicago Press).

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1-26.

Gruber, M., Gelman, B., & Ranganath, C. (2014). States of curiosity modulate hippocampus- dependent learning via the dopaminergic circuit. *Neuron*, 84, 486-496.

Hamlin, J. (2013). Moral Judgment and Action in Preverbal Infants and Toddlers Evidence for an Innate Moral Core. *Current Directions in Psychological Science*. 22. 186-193.

Harenski CL, Thornton DM, Harenski KA, Decety J, Kiehl KA. (2012). Increased frontotemporal activation during pain observation in sexual sadism: preliminary findings. *Arch General Psychiatry*. Mar; 69(3): 283-92.

Hatfield, Elaine; Cacioppo, John T., Rapson, Richard L. (1993). Emotional contagion. *Current Directions in Psychological Science*. 2 (3): 96–99.

Hennenlotter A, Dresel C, Castrop F, Ceballos Baumann AO, Wohlschläger AM, Haslinger B. (2009). The link between facial feedback and neural activity within central circuitries of emotion—new insights from Botulinum Toxin—induced denervation of frown muscles. Cerebral Cortex.19:537–542.

Hess et al. (2011). Convergent and Divergent Responses to Emotional Displays of Ingroup and Outgroup. *Emotion*, 11 (2): 286-298.

Hickok, G., (2008). Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. *Journal of Cognitive Neuroscience*, 21: 1229–1243.

—— (2014). *The Myth of Mirror Neurons: The Real Neuroscience of Cognition and Communication*, New York: W.W. Norton and Company.

Hoehl, Stefanie & Hellmer, Kahl & Hedqvist, Maria & Gredebäck, Gustaf. (2017). Itsy Bitsy Spider…: Infants React with Increased Arousal to Spiders and Snakes. *Frontiers in Psychology*. 8.10.3389

Jabbi, M., Bastiaansen, J., Keysers C. (2008). 'A Common Anterior Insula Representation of Disgust Observation, Experience and Imagination Shows Divergent Functional Connectivity Pathways.' PLOS ONE 3(8).

Iacoboni M., Carr L Dubeau MC., Mazziotta JC., Lenzi GL. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proc National Academy of Science USA*. 100(9): 5497-502.

Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, et al. (2005) Grasping the intentions of others with one's own mirror neuron system. PLoS Biol. Epub 2005 Feb 22.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning, *Journal of Experimental Psychology: General*, 110, 306–340.

Jackson, P.L. et al. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *Neuroimage* 24, 771–779.

James, W., (1884). What is an Emotion? *Mind*, 9:188–205.

Joyce, R. (2000). Rational Fear of Monsters. *British Journal of Aesthetics* 40, 209-24

Keysers, Christian. (2010). Mirror Neurons. *Current Biology* 19 (21): 971–973.

Kidd, C. & Hayden, B. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88, 449-460.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind.' *Science*, 330, 932.

Knoblich G., Butterfill S., Sebanz N. (2011). Psychological research on joint action: theory and data, in *The Psychology of Learning and Motivation*, eds Ross B., editor. Burlington, VT: Academic Press; 59–101.

Krupenye, C, Call, J. (2019). Theory of mind in animals: Current and future directions. *WIREs Cognitive Sci*ence. 10:e1503.

Lamarque, P. (1981). How Can We Fear and Pity Fictions?' *British Journal of Aesthetics* 21.4, 291-304.

Lazarus, Richard S. (1991). *Emotion and Adaptation*, New York: Oxford University Press.

Leknes, S., Berna, C., Lee, M.C., Snyder, G.D., Biele, G., Tracey, I. (2013). The importance of context: when relative relief renders pain pleasant. *Pain.* 154(3): 402-10.

Leslie, K.R., Johnson-Frey, S.H., Grafton, S.T. (2004). Functional imaging of face and hand imitation: Towards a motor theory of empathy. *Neuroimage* 21, 601–607.

Levy, D. J., & Glimcher, P. W. (2012). 'The root of all value: a neural common currency for choice.' *Current opinion in neurobiology*, *22*(6), 1027–1038.

Li, W., Moallem, I., Paller, K. A., & Gottfried, J. A. (2007). Subliminal smells can guide social preferences. *Psychological Science*, *18*(12), 1044-1049.

Liew, T. W., & Tan, S. M. (2016). The effects of positive and negative mood on cognition and motivation in multimedia learning environments. *Educational Technology & Society*, 19 (2), 104–115.

Mathews A, MacLeod C. (2005) Cognitive vulnerability to emotional disorders. *Annual Review of Clinical Psychology* 1: 167–95.

Martin, G. B., & Clark, R. D. (1987). Distress crying in neonates: Species and peer specificity. *Developmental Psychology*,18. 3-9.

Martens, WHJ. 2000. The hidden suffering of the psychopath. *Psychiatric Times*; 19(1): 1-7.

Martindale, C., & Moore, K. (1988). Priming, prototypicality, and preference. *Journal of Experimental Psychology: Human Perception and Performance, 14*(4), 661-670.

Meskin, A., & Weinberg, J. M. (2003). Emotions, Fiction, and Cognitive Architecture. The British Journal of Aesthetics, 43(1), 18-34.

McWhinnie, H. J. (1968). A review of research on aesthetic measure. *Acta Psychologica*, 28, 363-375.

Metcalfe, J. and Michel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, 106, 3-9.

Molenberghs P, Cunnington R, Mattingley J. (2009). Is the mirror neuron system involved in imitation? A short review and meta-analysis. *Neuroscience & Biobehavioral Reviews* 33 (1): 975–980.

Moran, R. (1994). The Expression of Feeling in Imagination. *Philosophical Review* 103.1: 75-106.

Neu, J. (2000). *A Tear is an Intellectual Thing: the Meaning of Emotions*. (Oxford, New*York: Oxford University Press).*

Neumann D.L., Boyle G.J., Chan R. (2013). Empathy towards individuals of the same and different ethnicity when depicted in negative and positive contexts. *Personality and Individual Differences.* 55(1): 8-13.

Nico H. F. (1986) *The Emotions*. Cambridge (UK): Cambridge University Press, 207.

Nichols, S. (2004). Imagining and Believing: The Promise of a Single Code. Journal of Aesthetics and Art Criticism, 62, 129-139.

--- (2004). Review: Recreative Minds. Mind, 113(450), 329-334.

Nussbaum, M C., *(2001). Upheavals of Thought: The Intelligence of Emotions*. (Cambridge: Cambridge University Press).

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. (New York: Oxford University Press).

Paskow, A (2004). *The Paradoxes of Art: A phenomenological investigation*. (Cambridge: Cambridge University Press)

Perner J., Frith U., Leslie A. M., Leekam S. R. (1989). Exploration of the autistic child's theory of mind: knowledge, belief, and communication. *Child Development.* 60, 689–700.

Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). A review and meta-analysis. *Neurophysiolical Clin* 30: 263–288

Phan, K et al. (2001) 'Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMR.' *NeuroImage*. 16(2):331-48

Prather, Jonathan & Mooney, Richard. (2014). *Mirror Neurons in the Songbird Brain: A Neural Interface for Learned Vocal Communication*.

Premack, D. G.; Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences* 1 (4): 515–526.

Prinz, Jesse. (2011) Against Empathy. *The Southern Journal of Philosophy*, 49 (Spindel Supplement): 214–233.

Radford, C. (1975) 'How Can We Be Moved by the Fate of Anna Karenina?' *Proceedings of the Aristotelian Society*, Supplemental Vol. 49, pp. 67-80.

— . (1977) 'Tears and Fiction.' *Philosophy* 52, pp. 208-213.

— . (1982a) 'Stuffed Tigers: A Reply to H. O. Mounce: Discussion.' *Philosophy* 57 (222): 529-532.

— . (1982b). 'Philosophers and Their Monstrous Thoughts.' *British Journal of Aesthetics* 22 (3):261-263.

Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., Bushnell, M.C. (1997). Pain affect encodedin human anterior cingulate but not somatosensory cortex. *Science* 277 (5328), 968-971.

Ramirez, I. (1990). Why do sugars taste good? *Neuroscience. Bio-behavioral Review.* 14: 125 –134.

Range F, Viranyi Z, Huber L. (2007) Selective imitation in domestic dogs. *Current Biology*. 17: 868–872.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition,* 8, 338–342.

Reber, R; Schwarz, N.; Winkielman, P. (2004). "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?". *Personality and Social Psychology Review*, 8 (4): 364–382.

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*, 45–48.

Reber, R. (2012). Processing Fluency, Aesthetic Pleasure, and Culturally Shared Taste. In Shimamura, A. P., & Palmer, S. E. (Eds.). (2012). *Aesthetic science: Connecting minds, brains, and experience*. New York, NY, US: Oxford University Press.

Reber, R., Wurtz, P., & Zimmermann, T. D. (2004). Exploring "fringe" consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition*, 13, 47–60.

Rizzolatti, Giacomo; Craighero, Laila (2004). "The mirror-neuron system" (PDF). *Annual Review of Neuroscience.* 27 (1): 169–192.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*. 39 (6):1161-1178

Russell, J. A., Barrett, L. F. (1999). 'Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.' *Journal of Personality and Social Psychology,* 76(5): 805-819

Ruse, M. & Wilson E. (1986). 'Moral Philosophy as Applied Science: A Darwinian Approach to the Foundations of Ethics.' *Philosophy,* 61 (236):173-192

Sagi, A., & Hoffman, M. L. (1976). Empathic distress in the newborn. *Developmental Psychology*,12: 175-176.

Sawa, K. (2009). Predictive behavior and causal learning in animals and humans. *Japanese Psychological Research*, 51 (3): 222–233.

Scherer, Klaus R., Angela Schorr, and Tom Johnstone (eds.), (2001) *Appraisal Processes in Emotion: Theory, Methods, Research*, (*Series in Affective Science*), Oxford: Oxford University Press.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as Embodied Moral Judgment. *Personality and Social Psychology Bulletin, 34*(8):1096-1109.

Schroeder, T., & Matheson, C. (2006). Imagination and Emotion. In S. Nichols (Ed.), The Architecture of the Imagination (pp. 19-40). Oxford: Oxford University Press.

Schwarz, N. (2010). Feelings-as-Information Theory. In P. Van Lange, A. Kruglanski, & E. T. Higgins (eds.), *Handbook of theories of social psychology*. Sage.

Schwarz, N., & Clore, G. L. (1996). Feelings and phenomenal experiences. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (433-465). New York, NY, US: Guilford Press.

Shuman, V. and Scherer, K. R. (in press). Concept and the structure of emotion, in *Handbook of Emotions in Education*, eds R. Perkun and L. Linnenbrink-Garcia (London: Routledge).

Silvia, P. J. (2012). Human Emotions and Aesthetic Experience. In *Aesthetic Science: Connecting Minds, Brains, and Experience*, edited by Arthur P. Shimamura and Stephen E. Palmer, (250-275). New York: Oxford University Press

--- (2006). *Exploring the psychology of interest*. New York, NY, US: Oxford University Press.

Smith, C. A., Kirby, L. D. (2000). Consequences Require Antecedents: Toward a Process Model of Emotion Elicitation. In J. P. Forgas (Sd). *Feeling and Thinking: The role of affect in social cognition.* 83 - 106. New York: Cambridge University Press.

Solomon, R. (1988) On Emotions as Judgments. *American Philosophical Quarterly*. (25)2:183-191.

Sperber, D., Wilson, D. (2002). *Pragmatics, Modularity and Mind-reading*. Mind and Language, Wiley, 17 (1), 3-33.

Steiner, J. E., Glaser, D., Hawilo, M. E., and Berridge, K. C. (2001). Comparative expression of hedonic impact: affective reactions to taste by human infants and other primates. *Neuroscience. Bio-behavioral Review* 25, 53–74.

Thompson, D. V., & Ince, E. C. (2013). When disfluency signals competence: The effect of processing difficulty on perceptions of service agents. *Journal of Marketing Research*, 50, 228–240.

Thomsen L. (2020). The developmental origins of social hierarchy: how infants and young children mentally represent and respond to power and status. *Current Opinion in Psychology*. 201-208.

Tooby, J., & Cosmides, L. (2008). The Evolutionary Psychology of the Emotions and their Relationship to Internal Regulatory Variables. In M. Lewis, J. Haviland, & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 114-137). (New York: Guilford Press).

Venables N.C., Hall J.R. Patrick C.J. (2013). Differentiating psychopathy from antisocial personality disorder: a triarchic model perspective. Psychological Medicine. 9: 1-9.

Viinikainen et al. (2010). Nonlinear Relationship Between Emotional Valence and Brain Activity: Evidence of Separate Negative and Positive Valence. *Human Brain Mapping.*

Vitz, P. C. (1966). Affect as a function of stimulus variation. *Journal of Experimental Psychology*, *71*, 74–79.

Wager, T. (2002). 'Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI.' *NeuroImage.* 16 (2): 331–48.

Walton, K. (1970). Categories of art. *Philosophical Review* 79 (3):334-367.

---. (1978). Fearing Fictions. *Journal of Philosophy* 75 (1):5-27.

Wallentin, M. et al. (2011) 'Amygdala and heart rate variability responses from listening to emotionally intense parts of a story.' *NeuroImage*. 58. 963–973.

Weinberg, J., & Meskin, A. (2006). Puzzling Over the Imagination: Philosophical Problems, Architectural Solutions. In S. Nichols (Ed.), The Architecture of Imagination (pp. 175-204). Oxford: Oxford University Press.

Wellman HM, Cross D, Watson J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72: 655–684.

Whittlesea, B., Jacoby, L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language*, 29, 716–732.

--- *(1993). Illusions of familiarity. Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1235–1253.

Wicker, B. et al. (2003). Both of us disgusted in "My insula: the common neural basis of seeing and feeling disgust." *Neuron 40*, 655–664.

Winkielman, P., & Fazendeiro, T. A. (2001). The role of conceptual fluency in preference and memory. Unpublished manuscript.

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17*, 799–806.

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (189–217).

Wurtz, P., Reber, R., & Zimmermann, T. D. (2008). The feeling of fluent perception: A single experience from multiple asynchronous sources. *Consciousness and Cognition*, 17, 171–184.

Zachar, Peter & D. Ellis, Ralph. (2012*). Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell*. (Amsterdam: John Benjamin's Publishing Company).